Project Final Report

Project Title:	**Helios**
Project Period:	**04/01/2013 – 09/30/2016**
Reporting Period:	**01/01/2016 – 03/31/2016**
Reporting Frequency:	**One time (final report)**
Submission Date:	**12/02/2016**
Recipient:	**SRI International**
Recipient DUNS #:	**009232752**
Address:	**333 Ravenswood Ave** **Menlo Park, CA 94205-3493**
Award Number:	**DE-EE0006130**
Awarding Agency:	**DOE EERE SEEDS Program**
Working Partners:	**University of Toledo**
Cost-Sharing Partners:	**None**
Principal Investigator:	**Lucien Randazzese** **Director** **Phone: 703-247-8628** **Fax: 703-247-8410** **Email: lucien.randazzese@sri.com**
Submitted by: **(if other than PI)**	**N/A**
DOE Contracting Officer:	**Diana Bobo**
DOE Project Manager:	**Christine Bing**

Executive Summary

This proof-of-concept project focused on developing, testing, and validating a range of bibliometric, text analytic, and machine-learning based methods to explore the evolution of three photovoltaic (PV) technologies: Cadmium Telluride (CdTe), Dye-Sensitized solar cells (DSSC), and Multi-junction solar cells. The analytical approach to the work was inspired by previous work by the same team to measure and predict the scientific prominence of terms and entities within specific research domains. The goal was to create tools that could assist domain-knowledgeable analysts in investigating the history and path of technological developments in general, with a focus on analyzing step-function changes in performance, or “breakthroughs,” in particular. The text-analytics platform developed during this project was dubbed Helios.

The project relied on computational methods for analyzing large corpora of technical documents. For this project we ingested technical documents from the following sources into Helios: Thomson Scientific Web of Science (papers), the U.S. Patent & Trademark Office (patents), the U.S. Department of Energy (technical documents), the U.S. National Science Foundation (project funding summaries), and a hand curated set of full-text documents from Thomson Scientific and other sources.

Topic modelling on these document groups highlighted the emergence of various technical approaches within each field, the replacement of one approach with another approach, and a preliminary mechanism by which we could use the Helios platform to automatically identify instances of this topic replacement. A human analyst could use Helios to review a much larger number of documents than he or she could ever read. This machine-assisted insight could point to important technical trends in specific domains that would help inform policy and further investment decisions.

We also developed an initial methodology for investigating how topic models evolve over time, and for using external technical libraries (in this case Wikipedia) to find user-friendly and human-meaningful names for topics in an automated fashion.

Finally, we developed a preliminary approach to aligning disparate documents groups in the analysis of the same technology evolution. Document types, such as scientific papers and patents, tend to have widely different structure and patterns of language and so cannot typically be used commonly to build topic models. By clustering papers and patents into clusters of topically similar documents, we can align clusters of papers to clusters of patents based on the terms common to both clusters. Our alignment approach is preliminary and requires refinement, but shows promise as a means to use disparate documents groups to analyses specific developments.

Table of Contents

Background

This project draws on two different streams of research: the development of text analytics and computational methods for analyzing technical documents on research efforts; and investigations into the nature of innovation dynamics, especially using bibliometric analysis. Recently, the convergence of these two research streams has accelerated, with more scholars exploiting techniques such as topic modeling and semantic analysis in conjunction with bibliometric analysis. The HELIOS project incorporates and aggregates key concepts from both research streams and applies these concepts over a wider pool of documents.

In text analytics, identifying networks and clusters of authors, terms, and ideas has been proven a crucial element of identifying trends and important events in technological evolution. McKeown et al. (2016) describe the results of their work under the IARPA FUSE program to determine how to predict the impact of newly emergent technical concepts on future science and technology. This experiment is especially significant, as they find that the analysis of full-text article data improves prediction substantially over the use of metadata-only publication records. HELIOS analyzes both metadata-only (WoS) and full-text publications.

A working paper by Bettancourt and Kaiser (2015) from the Santa Fe Institute, based on Bettancourt's earlier work studying network densification in scientific communities, finds evidence of a unique pattern in the evolution and emergence of a new scientific field. The authors extracted works on a collection of distinct topics and measured the growth in the number of authors and co-authorship linkages, for each field over time. They determine that for these fields, the networks of co-authorship grow more densely interconnected until they reach a "tipping point," after which these networks move to a growth path that is discontinuous with growth in prior periods. This change in the topology of networks appears indicative of the emergence of a distinctive new field of science. Small et al. (2014) built two large scale models of the overall scientific literature: one based on direct citation and the other based on co-citation. By combining these models, they were able to nominate emerging topics and identify clusters of new and rapidly growing technologies. They searched the press for recent major awards associated with the found topics, and were able to demonstrate a correlation. With Helios, these two methodologies have, to some degree, been joined. Helios is capable of establishing a co-authorship network surrounding a breakthrough, analyzing citation and co-citation patterns, but improves upon these bibliometric approaches by clustering concepts using advanced topic modeling techniques.

The team led by Varun Rai at the University of Texas at Austin (also a SEEDS award recipient) has explored the computational analysis of patent text. In Rai et al. (2013), they examined patent activity in four categories of the distributed solar PV balance of system (BOS) technology—inverters, mounting systems, monitoring, and site assessment—using patent data from a unique dataset constructed through keyword searches of the claims section of patents and patent applications in USPTO. Using the patent database from Rai et al. (2013), Venugopalan and Rai (2015) used topic modeling to map the patents to probability distributions of the real world BOS categories, demonstrating that linguistic features from topic models can be used to effectively identify the main technology area that applies to a patent's invention. The authors used topic distributions as features to train a technology-area classifier and then used several classification algorithms, such as linear discriminant analysis, quadratic discriminant analysis, neural networks, and SVM to distinguish between the four BOS technology areas – these automated classifications are then checked against patents that were manually classified in Rai et al. (2013). This supervised, automated approach is very similar to the Helios project and adds new analytical capabilities that are often difficult to achieve using conventional manual methods, where the time and resources required to identify and classify relevant patents in large databases is unduly burdensome. Ultimately, both Helios and the work by Rai et al. demonstrate that computational and automated analytical techniques are most effective when paired with human experts.

Recent advances are being produced by research teams funded by the Intelligence Advanced Research Projects Activity (IARPA) Foresight and Understanding from Scientific Exposition (FUSE) program (including researchers who are partners in SRI's Helios project for FUSE, on which Helios is based). Prof. Dan Roth and his colleagues at the University of Illinois has developed an approach to identify and characterize scientific concepts contained in scholarly publications (Roth et al. 2013). This approach opens up new opportunities to identify specific developments within a domain, and see how the researchers in that domain advance knowledge and practice around technologies. As Roth et al. state:

> if we want to achieve a deeper understanding of a scientific community from the paper trace generated by the community, there is a need to better analyze the text itself; there is a need to identify mentions of scientific concepts, categorize them and cluster them into coherent concepts, and study the relations between concepts of various categories.

In Helios, we use this approach to identify particular concepts described in the literature and salient to advances in PV technology, and how they relate to one another. A somewhat comparable approach is taken by Packalen and Bhattacharya (2012), in the patent literature. Rather than using natural language processing techniques, this experiment identified labeled innovations which became inputs to downstream progress. In developing Helios, we posit that novel combinations of different concepts are likely to be associated with particular breakthroughs. We use Helios to track the evolution of ideas and demonstrate how computational techniques can be used to monitor the convergence of concepts prior to a breakthrough.

We also leverage additional work done in the field of scientometrics, where scientific and technological advances are traced using metadata extracted from datasets of documents. A large number of papers is being produced by research institutes and academic institutions in East Asia, particularly the People's Republic of China and South Korea. For example, Luan et al. (2013) study how the diversification and convergence in solar technology inventions and related technologies coincide with progress in the solar energy technology sector. Helios combines advances in metadata analysis with full-text analytics to examine technological breakthroughs from multiple dimensions.

Many of the investigations into the nature of innovation dynamics use patent data and involve attempts to categorize and classify types of patents. Strumsky and Lobo (2015), also SEEDS grant awardees, report on their analysis of the U.S. patent database to investigate patterns in recombinant innovation. In this study, they use patent classification codes as a proxy for technologies. Any patent that is classified using multiple distinct and disparate codes is, in a sense, a combination of those technologies. They define four types of combinations: origination, where all codes that co-occur in a patent are new to the database; novel combination, where at least one of the codes that co-occur in a patent is new; combination, where all codes that co-occur are pre-existing, but did not previously co-occur; and refinement, where the codes and their co-occurrence have appeared in prior patents. Consistent with the theory, they find that instances of origination are uncommon and increasingly rare, while novel combination is also rare but somewhat more prevalent.

Leydesdorff et al. (2014) use patent analysis and patent maps to provide an analytical lens for studying the social networks of (co-)inventors, the geography of inventions, and the patterns in the knowledge bases of inventions. The authors use patent attributes such as inventor addresses, Cooperative Patent Classifications (CPC), and backward and forward citations as inputs for two different patent mapping programs with interactive overlays, allowing for dynamic animations and side-by-side comparisons. The authors demonstrate this technique on patents related to CuInSe2 thin film solar cell technology, and making the historical narrative of that technology's development evidence-based by showing the geographic footprints of the technology's development over time. We perform a similar analysis to study the evolution of ideas and breakthroughs in three solar photovoltaic technologies: dye-sensitized solar cells (DSSCs), multijunction solar cells, and cadium-telluride (CdTe) solar cells.

Zhou et al. (2014) profile research activity trends for DSSCs by applying two patent analysis methods: using patent family member classification information to show research activity trends and market shift processes, and using International Patent Classifications (IPCs) to trace technology development. This approach was used to try to answer three questions: how can research activity trends be estimated; how can market expansion patterns be identified; and what are the likely evolutionary paths of development? The authors found that DSSCs are in a rapid growth stage, various R&D organizations are identifying their promising commercial value, and that because DSSCs have good prospects for development, it is crucial to identify the main subsystems and the evolutionary paths for key topics in order to guide R&D management. While this paper focused on patent data, the authors noted that it would be possible to weave in analysis of scientific publications to enrich understanding of future DSSCs prospects.

Existing research has focused on patent data because of the rich information that can be derived from patent classification codes and linear citation trajectories. Helios analysis was predominately performed on World of Science (WoS) data and demonstrates that many of the methodologies and ideas that were used to analyze patent data are effective when applied to different document groups, especially when computational power can be leveraged to analyze metadata and full-text publications.

Helios was especially useful at analyzing different types of documents and connecting types of scientific publications together.

In a conference paper, Zhang et al. (2015) describe a quantitative method for linking basic laboratory research to patenting through the calculation of statistics on patent data and non-patent citations (NPC) found in patents. The analysis measures “technology coupling,” analogous to our idea of conceptual integration, to show how new inventions are produced by recombining existing sub-technologies. While the analysis shows a positive relationship between basic research activity and invention, it should be noted that this conclusion was developed based on an analysis of chemical patenting, and may not apply across domains.

In Helios we not only create clusters of terms using patent documents, but also cluster WoS terms and attempt to pair these two sets of terms together to match scientific research with resulting patents. The work completed in HELIOS demonstrates that there is much more to be done in this field but illustrates the ways that computational power can be leveraged to analyze breakthroughs across many dimensions.

Introduction

The Solar Energy Evolution and Diffusion Studies (SEEDS) program seeks to improve our understanding of solar technology evolution, ultimately providing methodologies to design policy interventions that can accelerate technological development and lead to rapid breakthroughs. Traditional approaches to describing technology evolution are either too simplistic (e.g., learning curves) to inform the policy making process or too slow and resource intensive (e.g., micro-level qualitative case histories) to provide timely advice.

Technological improvement in solar energy technology has not followed a smooth curve, but instead is characterized by unpredictable leaps in capability followed by periods of incremental advance. We can view these leaps as significant “breakthroughs” which either overcome long-standing technical “bottlenecks” to achieve greater improvements in performance relative to cost, or created wholly new approaches for achieving specific technical performance levels such as solar cell conversion efficiency.

Understanding and modeling these patterns of punctuated equilibria require both a new analytical framework, and new tools to capture the scope and nature of solar energy research and development. Such tools are especially helpful in synthetic fields such as solar technology, where critical breakthroughs often demand the integration of multiple and diverse knowledge domains such as materials science, nanotechnology, biology, and economics.

Project Objective

The objective of this project was to build and validate a proof-of-concept computational and analytical tool that could identify and describe the technical evolution of solar technology in a much shorter time period and with greater scope and more precision than was achievable in the past. The science-based tool was developed to provide data-driven insights on past patterns in research and development of solar energy systems to inform future decision-making by stakeholders aiming to make solar energy cost competitive with other forms of energy generation. Advantages of this approach include, potentially and for example, the ability to develop case studies of technology evolution more quickly, to automatically identify technology developments from a very large and diverse body of technical documentation that no one human analysts could examine, and to identify the sequence of research paths that led to specific breakthroughs in technology performance.

No human analyst or even team of analysts could ever review the thousands-to-millions of documents our machine-based approach can. This capability is thus most obviously useful in organizations with broad missions and thus for which deep expertise is limited. When studying any given technology, all they could aspire to do well is a limited, science literature and popular press-driven review of that technology's history and state of art. To be able to run a semi-automated analysis that identified and quantify key performance improvements in virtually any technology area and the papers, scientists, institutions, and topics of inquiry associated with those breakthroughs would be of great use.

Even within organizations with deeper expertise, however, we believe the methods developed during this project would be of use. Most experts in a given technology are have a particular view of the history of their technology that highlights those aspects most aligned with their own training and experience, and most (if not all) have strongly biased views about which approaches make the most sense for research going forward. Helios can be used in this case to offer alternative suggestions about the interrelation between research threads in the field and can be used to help organize central contributors to those areas that only touch peripherally on the knowledge of experts.

Also, for most fields, there are simply too many new publications for any reasonably sized group of experts to keep up with, especially as research become increasingly interdisciplinary. In many fields, such as the biological sciences, there are an enormous number of papers that each hold a very small clue about how some highly complex system works, and all of the information is valuable. In this case, Helios can help a group of experts identify the most significant new publications and recognize influential papers early.

Some fields feature a relatively small number of known main contributors, and it may, in principle be sufficient merely to keep up with publications from these researchers alone. However, this creates obvious potential bias in future research the leads attention away from newly emerging ideas and newly emerging scientists, at a potentially great cost to research impact.

The tool to be developed and tested is based on adapting new computational approaches developed for the study of technology emergence to describe and provide insight into the evolution of solar energy technologies. The overall approach is based on the availability of large digital repositories of documents such as scientific publications, patents, funding records, and other relevant sources, combined with advanced machine learning-based computational techniques for the systematic identification and analysis of developments in solar technology.

The project used a series of human-created case studies of various solar technologies – e.g., crystalline silicon, cadmium telluride (CdTe), etc. – as the “ground truth” to which the machine-based results are compared and validated.

Project Team

This project is conducted entirely by staff from SRI International, a non-profit research corporation based in Menlo Park.

- Technology Analysis Group. Staff from the Center for Innovation Strategy and Policy (at the start of the project, the Center for Science, Technology and Economic Development) at SRI’s Washington, DC office, led by co-PI Jeffrey Alexander. Formulated research questions and study models, conducted technology case histories, identified key indicators and validated Helios performance in entity tracking.
- Advanced Analytics Group. Staff from the Artificial Intelligence Center, based in SRI’s San Diego, California office, developed algorithms, software, data infrastructure, and tools for applying sophisticated text analytics to datasets of technical documents numbering from thousands to millions of records.

SRI obtained expert input from internal and external advisers.

- Current and emeritus laboratory directors from the SRI International Materials Research Laboratory in Menlo Park, with extensive technical experience in evaluation of PV materials.
- Lead research staff from the Wright Center for Photovoltaics Innovation and Commercialization at the University of Toledo, Ohio, with special expertise in CdTe thin film technology.
- Scientific leadership and research staff for the Solar Portfolio at General Electric Global Research, Schenectady, New York

Technical Work Plan

The work plan for the project extended from April, 1, 2013 to September 30, 2016, with a three-month grace period to complete final project reporting. This period was divided into two periods of performance with a single go/no-go decision after period one which ended on June 30, 2014. The original Statement of Project Objectives (SOPO) for the entire project, including relevant tasks and milestones, is described below.

Budget Period 1 (15 months)

In the first period, the project team will work closely with our expert advisory panel to refine the research design and develop the software tools that will serve as a solid foundation for execution in period 2. We will conduct concurrent research on solar technology evolution as a means of validating the outputs generated by our computational platform. By the second half of period 1, preliminary versions of the tool running on SRI’s in-house system will be ready to support exploratory analysis of a variety of solar innovations, as data becomes available for processing.

Task 1: Research Design

How can the Helios platform enhance our understanding of solar technology evolution, and more importantly, how can its outputs help industry to develop and commercialize solar innovations? Task 1 will focus on answering these questions through intensive interaction between the SRI project team and the expert advisory panel. The results from Task 1 will inform the design of activities in all subsequent tasks.

- 1.1: Refine research questions. In subtask 1.1, SRI will work with the expert advisory panel to identify the information used by solar technology practitioners and policy makers to guide program management and funding decisions, and refine our research questions to respond to those needs. The research questions will guide our investigations of the conditions around past improvements in solar technologies in an attempt to inform future actions that may accelerate the evolution of solar technology. We will focus in particular in the process for connecting advances in research and development with market outcomes, such as price reductions. For example, we propose to explore the relationship between the co-evolution of competing solar technologies and improvements in the efficiency of solar conversion processes, and will work with panel members to identify likely technology “targets” that influenced those improvements.
- 1.2: Identify and evaluate suitability of data sources. Traditional data sources such as peer-review journal articles and patents are a key data source for the proposed approach; however, these data sources suffer from significant limitations such as lag or gaps in coverage on key topics. In tandem with the advisory panel, SRI will explore alternative data sources that may provide insights into market and regulatory signals which may affect technology development decisions. Some of these new sources may be appropriate for automated analysis using the Helios platform while others may require manual analysis.
- 1.3: Develop dissemination plan. Identification of appropriate forums and dissemination mechanisms to maximize the impact of the proposed research (focusing on peer-reviewed journals and conferences with academic and industrial audiences). Drawing on the advisory panel’s industry and technical expertise, SRI will develop a plan to disseminate its results to the widest possible audience.
- Task 1 Milestone: Development of revised project plan incorporating advisory panel input.

Task 2: Development of Baseline Case Studies for Three Solar Technology Breakthroughs

The project team will develop case studies detailing the evolution of three carefully chosen solar innovations. Solar cell and module technology-driven innovations will be the primary focus of Helios-driven case studies; however, the team will explore a broader scope of solar innovation, potentially including complementary technologies or processes if sufficient data sources can be identified and acquired. Technologies will be chosen to enable a detailed cross-case comparison that distinguishes between technical barriers and dynamics which are commonly seen across a range of technologies, and those that are unique to a particular technical approach. Relevant subtasks and milestones include.

- Subtask 2.1: Select appropriate targets for detailed analysis. A list of potential technologies (crystalline silicon, CIGS, CdTe, dye-sensitized solar cells (DSSC), etc.) and selection criteria will be developed and discussed with the advisory panel. Based on this input, three solar technologies will be selected for detailed analysis. Selection criteria will include: availability of relevant data sources, technical expertise of the advisory panel, and potential for the case to provide useful insights into the process of technology evolution.
- Subtask 2.2: Develop baseline case studies. The SRI team will develop baseline case studies employing primarily qualitative methods, but also incorporating quantitative methods such as bibliometric and scientometric analysis. These case studies will play an important role as baseline, or ground-truth, studies to validate later quantitative results produced through the Helios platform or through exploratory analysis of alternative data sources.
- Subtask 2.3: Case study validation. Case studies will be submitted to the expert advisory panel for review. If the experts on the advisory panel feel that they are not equipped to evaluate the output for a specific case study, additional expertise will be identified as necessary to provide appropriate feedback and guidance. Supplemental expert reviewers will be identified both by referrals from panel members and from bibliometric research to identify central researchers involved in the particular technology in question.
- Subtask 2.4: Preliminary cross-case analysis. SRI will conduct a cross case analysis to identify key similarities and unique features of the evolutionary paths of solar technologies. The results of this analysis will inform that development of the Helios platform by identifying the classes of features and relationships between these features that characterize solar technology evolution.
- Task 2 Milestone: Completion of three detailed case studies to serve as baseline for later efforts and one peer-reviewed publication, which reviews the major features of our case studies and ties these features into a scientific model of technology evolution.

Task 3: Helios Platform Development

The Helios platform is derived from the Helios software platform that was designed and constructed under the Foresight and Understanding from Scientific Exposition program to detect scientific emergence, sponsored by the Intelligence Advanced Research Projects Activity. We will configure, customize, and apply Helios to data on the evolution and diffusion of solar energy technologies. This tool (termed Helios) will support the Technology Analysis team by directing their research through the available literature.

- Subtask 3.1: Data acquisition. We will begin with sources identified in task 1. We will set up computing and storage infrastructure to house these data locally, and will arrange for obtaining the data. Some of these will be simple downloads, some may require negotiation with the data owners, and some will require web scraping.

- Subtask 3.2: Data ingest and disambiguation. Each of the sources collected will need to be parsed and stored in the Helios database. This provides efficient, source-independent access to text and metadata for downstream processing. Helios disambiguation techniques will be customized for the obtained data sets. These disambiguation techniques identify and merge duplicate references to a single entity. (For example, identify that “JOHN A. SMITH” and “J.A. SMITH” are the same person.)
- Subtask 3.3: Topic Modeling. We will apply information-theoretic co-clustering of terms and documents to identify term clusters that pick out coherent topics. Documents will be classified based on their distributions over topics. The Helios interface supplies a browser that will allow the analysis team to look at the relations between topics, to identify groups of documents that distribute over documents of interest, and to examine individual documents of interest.
- Subtask 3.4: Entity Tracking. We will track the careers of specific authors and inventors over time, both through the space of topics on which they write and through the space of organizations with which they affiliate (the assignees of their patents and the affiliations on their publications). We will also track the affiliated organizations themselves. Again, this will allow the analysis team to quickly identify key actors in order to examine their documents more closely.
- Subtask 3.5: Network Influence Analysis. Helios provides the ability to analyze a large set of patents or publications and determine the strength of influence that any given document has on later documents. This scoring will be used to nominate key patents and papers for examination by the analysis team. We will attempt to extend this analysis to estimate the influence of documents on downstream activities (e.g. whether a particular finding influenced later researchers to change their research priorities).
- Subtask 3.6: Analysis Team Support. The analytics team will dedicate time during the second half of the phase for guiding the analysis team through use of the software tools and for making minor customizations to the tools in order to improve usability. This will significantly increase the productivity of the analysis team.
- Task 3 Milestone: Helios will have output a co-clustering interface that the Technology Analysis team can use to identify key topics in the data and groups of documents that display certain patterns with respect to these topics. Helios will present influence metrics of key publications, based on analysis of citation graphs. Helios will produce initial results in entity tracking that will be used by the Technology Analysis team in their case studies.

Go/No-Go Decision Point

The team will deliver a preliminary set of results from the Helios system focused on the same topics as the human-generated case studies. By the end of phase 1, Helios will generate preliminary outputs such as topics, tracked entities, and influential documents that that can be assessed for congruence with the findings of the human-generated case studies. The assessment criteria at the end of phase 1 will illustrate the progress made toward the development of a system that identifies and characterizes the relationships between entities (people, organizations, and documents) and technical developments affecting the impact of the solar innovations studied. Three key dimensions for assessment include:

- Efficiency: The computational efficiency of Helios and the degree to which Helios accelerates discovery.
- Coverage: The scope of coverage and volume of documents and other data that Helios processes to generate those outputs.
- Insight: The degree to which Helios output is comparable to (or exceeds) the accuracy and salience to the findings of the human case studies.

These three dimensions are largely subjective, but the following proxy measures will illustrate progress at the end of phase 1 as described in Table 1 below.

Dimension	Indicator	Target
Coverage	Number of different data sources identified, ingested, and analyzed. (scientific literature databases, patents, technical report databases, etc.)	3-5 data sources
Insight	Helios identifies highly influential documents for each technology topic. We will take the case studies to provide ground truth for estimating recall (the percentage of actual high impact papers that were identified) and manual review to identify precision (the percentage of identified papers that were actually of high impact).	50% Precision; 75% Recall, or on par with human
Insight	Helios recognizes the main solar innovations and their components as illustrated in the case studies. (Evaluated as above.)	35% Precision; 75% Recall, or on par with human
Insight	Helios associates the same entities (institutions, authors, etc.) with each innovation as the case studies.	50% Precision; 50% Recall
Efficiency	Document processing speed.	100,000 documents in less than 24 hours.

Table 1: Progress Evaluation Criteria

Budget Period 2 (27 months)

In the second period, the focus will shift from design to execution. In period 2, the advanced analytic tools that were developed and tested in period 1 will be applied to the study of solar technology evolution. Period 2 will be characterized by frequent interaction between the Technology Analysis group and the Advanced Text Analytics group.

Task 4: Helios Platform Extension

The Helios platform will be extended based on our experiences in Period 1. We will extend it to include more temporal analysis and will implement features based on the needs of the Technology Analysis group.

- Subtask 4.1: Supplemental Data Collection. Phase 1 work will focus on the collection, ingest and processing of five separate, distinct datasets. We will move beyond the state-of-the-art by integrating these datasets, both in data management systems and in analytical techniques, so that we can achieve greater insight from comparisons across datasets, and enable cross-data validation of our findings.
- Subtask 4.2: Data ingest and disambiguation. These tasks will need to be applied to any new data acquired. We will also take the opportunity to improve the quality of the disambiguation achieved in phase 1, in order to provide data more easily interpretable by the analysis team.
- Subtask 4.3: Advanced Topic Modeling. We will extend the phase 1 models with temporal analysis of topics over time. We will associate topics each year with topics in the next year, based on term distributions. As originally planned, we intended to study birth, death, convergence, and divergence of topics. But more interesting now is the indication that we may be able to recognize when one topic is replacing another (like solid state over aqueous DSSC) or when basic science is starting to yield application. We will investigate these topic pairs using the case studies to identify features that nominate such transition. We expect citation structure to play a significant role in this investigation as well. We will also look at varying the granularity of topic models and the induction of hierarchical models in order to automatically group topics into families and split them into subtopics.
- Subtask 4.4: Advanced Entity Tracking. In phase 2, we will extend the phase 1 entity tracking model based on feedback from the technology analysis team. This may involve different treatment of specific types of entities, or may involve tracking entity relationships over time.
- Subtask 4.5: Additional Indicators. While using the tool and studying the solar technology literature, the analysis team will arrive at a variety of hypotheses that will require dedicated software support in order to explore. This task provides for this support.
- Subtask 4.6: Analysis Team Support. The analytics team will dedicate time during the second half of the phase for guiding the analysis team through use of the software tools and for making minor customizations to the tools in order to improve usability. This will significantly increase the productivity of the analysis team.
- Task 4 Milestone: Helios will be a stand-alone system that fully supports topic modeling, entity tracking, and network influence analysis in support of Technology Analysis over all obtained data sets.

Task 5: Platform Validation and Deployment

- Subtask 5.1: Perform in-depth Helios-aided exploration of case study topics. Using the expanded capabilities of the Helios platform, the team will identify key breakthroughs in the development of solar technologies and explore the environment surrounding those breakthroughs. The team will explore themes such as: team structures through network analytics, funding mechanisms through scientific literature metadata, and research approaches through classification of identified topics.

- Subtask 5.2: Report and Validate Results. Helios output will be incorporated into narratives describing the evolution of the target solar technologies. These narratives will be compared to the previously vetted human generated case studies to assess the validity of the results.
- Task 5 Milestone: Completion of three data-driven quantitative case studies incorporating findings from task 2 and quantitative results from the Helios platform.

Task 6: Method Expansion

- Subtask 6.1: Expansion beyond small set. The validated Helios platform will be used to describe the evolutions of solar technologies beyond the three technologies targeted in earlier tasks. In particular, we plan to investigate another emerging technology, such as perovskite solar cells, using the Helios platform, and then request that our expert panel and other subject matter experts evaluate the outputs to provide further validation and recommendations for system development.
- Task 6 Milestone: Completion of tests of broader applicability of the Helios platform.

Project Results and Discussion

Task 1: Research Design

Subtask 1.1: Refine Research Questions

The project team’s original research intension focused on using text analytics to investigate breakthroughs in photovoltaics (PV) technology over time, particularly on identifying preconditions to breakthroughs, such as funding and team structures; integration of prior research streams; and technology-specific trajectories. Feedback from the project panel of technology experts advised making research questions more specific. As a result, the project reoriented around proof-of-concept analyses such as:

- Can text analytics be used to identify, on a human-aided basis, emerging topics in solar development?
- Can the evolution of topics as they diverge into multiple separate topics or coalesce into a smaller number of topics be traced?
- Can the replacement of one technical approach by another be identified in an automated fashion?
- What type of performance indicators can be identified via text analytics?

The **Task 1 Milestone**, development of revised project plan incorporating advisory panel input, included these and related focus areas.

Subtask 1.2: Identify and Evaluate Suitability of Data Sources

In order to address the project research questions, five categories of text data sources were identified, as described in Table 2 below.

Source	**Description**
Thomson Scientific Web of Science	Scientific publications (primarily journal articles) in XML format identified through refined Boolean searches specific to each of the target technologies.
U.S. Patent & Trademark Office	All patents available in XML format. Collected all patents regardless of relevance to PV
U.S. Dept. of Energy	Technical reports and papers available through the Office of Scientific & Technical Information (OSTI) SciTechConnect website.
U.S. National Science Foundation	XML formatted grant records for funding awards, regardless of relevance to PV.
Thomson Scientific & others	Selected full-text versions of scientific publications, identified by human analysts as important documents for each target technology.

Table 2: Project Data Sets

The DOE Technical Reports and manually curated full-text (rows 4 and 5) are semi-structured or Unstructured data. Through its discussions with NREL, the SRI team identified the Solar Cell Efficiency Tables series of publications as a source of references to specific sources that make cell efficiency claims.

Subtask 1.3: Develop Dissemination Plan

The dissemination plan focused on journals and venues that cover results of text analysis based research of technical documents related to scientific and technological innovation. Conference publications were delivered at:

- Proceedings of the 2014 STI Conference on Science & Technology Indicators
- 2015 Atlanta Conference on Science & Innovation Policy

Conference presentations were made at:

- 2013 and 2015 Atlanta Conference on Science & Innovation Policy
- FEDLINK Symposium on Analytical Methods for Technology Forecasting
- 2014 SunShot Summit

A paper has also been submitted to *Technological Forecasting & Social Change*, and another paper on automated claims analysis is in progress; no outlet has yet been selected.

Task 2: Development of Baseline Case Studies for Three Solar Technology Breakthroughs

Subtask 2.1: Select Appropriate Targets for Detailed Analysis

In parallel to refining the text-analytic work of the project, the project team prepared detailed human-generated (manual) case studies, using the following criteria to select technologies for the manual case studies:

- Key developments in each technology occurred after the year 2000 (adequate digital documents available)
- The technologies together represent a diversity of technical approaches to PV
- The technologies range in their developmental stage from relatively mature to emerging
- Each technology is “interesting” for some reason, in the view of the advisory panel members

We selected for case analysis multi-junction cells (MJ), CdTe thin film cells, and dye-sensitized solar cells (DSSCs). MJ cells are notable for the very high efficiency rates achieved in the laboratory. CdTe represents a more mature technology with lower costs of manufacturing than most other technologies. DSSCs are an emerging technology.

Subtask 2.2: Develop baseline case studies

The Technology Analysis Group compiled case histories of the three target technologies using a combination of document research and interviews with experts knowledgeable about the development of each technology. The case histories were not intended to be research studies themselves, but to provide qualitative “ground truth” about the nature and circumstances of technical breakthroughs across a diverse set of technologies. They identified significant breakthroughs, concepts, researchers, institutions, seminal works, and similar entities. Case history write-ups ranged from 10-to-30 pages, excluding tables, figures and references, and each cited 25-to-75 source documents.

These case studies were exploited in two ways. First, they identify particular artifacts (e.g., scientific papers), entities (e.g., researchers), and events (e.g., discoveries) associated with particular breakthroughs in each of the three technology domains. Second, they provide a form of “ground truth” to determine to what extent the Helios platform can reconstruct a narrative case study of how innovation in each technology domain unfolded.

Subtask 2.3: Case Study Validation

During case history development, each of our human analysts spoke to experts from each technology group to verify facts and findings from document research. We also made extensive use of survey articles and trade press to complement technical publications reviewed.

Subtask 2.4: Preliminary Cross-Case Analysis

The case studies were prepared as ground truth against which to compare our machine-generated results. Nevertheless, there were some interesting conclusions drawn from across the set of three case studies.

- R&D Progress – To advance from a theoretical concept to a mature technology requires initial exploration of potential enabling technologies in the laboratory. However, advances in manufacturing of related materials often prompt such breakthroughs (glass manufacturing for CdTe, silicon wafer fabrication for MJ).
- Technology Commercialization – Some technologies are market ready upon leaving R&D labs. However, the vast majority of technologies require additional development from firms interested in commercial opportunities using the technology. We found that government agencies were a critical factor in this transition (NREL & NASA for MJ, Department of Energy for DSSCs).

The case studies also highlight a number of indicators of innovation:

- Sharp increases in volume of publications on that technology (indicates interest/acceptance)
- Development of “seminal works,” and involvement of prominent researchers (indicates legitimacy)
- Geographic spread of research (indicates acceptance and estimate of potential progress)
- Involvement of industrial researchers (indicates estimate of commercial potential)

These indicators correlate with bibliometric approaches to document analysis. Project machine analysis uses these indicators as a starting point, but goes well beyond such bibliometric approaches, as described below.

Task 3: Helios Platform Development

The Helios platform was based on the Helios platform developed under the IARPA FUSE program. The Helios platform was separated from Helios development in order to facilitate customization for the SEEDS program. The final software package includes the ability to run indicators from an early version of Helios, including Page Rank for citation, co-citation, and co-authorship networks. Helios utilizes a PostgreSQL database to store document metadata and the text of titles and abstracts. Helios algorithms are exposed through a REST API that allows users to interact with the database programmatically. The REST API can be accessed through a custom program (such as python or java), through a Swagger interface, or through curl on the command line. Helios is delivered to the Department of Energy in a Virtual Machine with all required database tools installed and with the database fully loaded. Custom scripts for generating the results reported on in this document are included in the Virtual Machine.

Subtask 3.1: Data Acquisition

We acquired and ingested into Helios data as summarized below in Table 3.

Source	**Documents Ingested**
Thomson Scientific Web of Science	Nearly 200,000 XML documents, plus over 2.3 million citations.
U.S. Patent & Trademark Office	4.9 million grants and applications, plus approx. 8 million citations.
U.S. Dept. of Energy	8,160 PDF files identified by searching for “solar” (5,583 technical reports and 2,577 conference papers)
U.S. National Science Foundation	Over 500,000 award records
Thomson Scientific & others	119 publications

Table 3: Project Data Sets – Document Coverage

We were able to utilize a computer server already present in the Artificial Intelligence Center for processing of this data. We procured a dedicated external hard disk drive for storage of all data. Negotiating data access, particularly from commercial sources, took longer than expected and delayed our research progress.

Subtask 3.2: Data Ingest and Disambiguation

We adapted parsers from Helios to ingest the above data sources into the Helios system. Helios is supported by a MongoDB database which stores objects representing documents, authors, institutions, and publication venues. New objects are created as raw documents are parsed, and new entries are generated for every citation as well as for every raw document record. Duplicate entries are then resolved.

PDF files were accompanied by metadata in structured format, which was ingested into the Helios database. We ran some initial tests for extracting plain text from PDF and we believe that the process will not be problematic, but we prioritized our Year 1 analyses on the metadata (which includes abstracts).

Subtask 3.3: Topic Modeling

We applied topic modeling from the Helios system to the three doc groups for multijunction, CdTe, and DSSC technologies. Topics are groups of occurrences of terms that are related by their co-occurrence in documents. The same term might belong to multiple topics if it co-occurs with different groups of terms. For example, the term “cell” might belong to a topic about cellular biology and might also belong to a topic about cellular telecommunication. The model assigns each term occurrence to a topic. Each occurrence belongs to a unique document, so we can look at the distribution of topics over time based on document publication dates.

We illustrate this temporal data for the case of DSSC in Figure 1 below, where rows are topics, columns are years, and the color of a cell indicates the number of occurrences of that topic in that year. 100 topics are shown (the selection of 100 as the number of topics is arbitrary). The reddest cells in Figure 1 indicate the largest number of occurrences of a topic in a given year, and the bluest cells indicate the smallest number of occurrences. Yellow cells are between blue and red. The brighter colors on the right half of the figure represent the increased number of publications on DSSCs in more recent years. The darker colors in the rightmost column indicate that our data for the most recent year is probably incomplete. This is a common occurrence in bibliometrics, as many documents are often added to a collection only a year or more after publication.

We experimented with a number of information-theoretic metrics for identifying the most pertinent terms to summarize a given topic, settling on pointwise mutual information (PMI) between term and topic. This is a measure of how specific a term is to a given topic, but also how much coverage the term has. The best term for topic is one with the optimal trade-off between frequency in the topic and specificity to the topic. Because topic labels are not an output of the topic modeling, we use this PMI "best term" as the topic label, which is presented to the right of each row in the figure, as a name for the topic (note that this means more than one topic can have the same name because individual terms can be included in more than one topic). This would allow a prospective analyst using Helios output of this type to get an idea of what each topic might be about.

We ran a stochastic search to determine an ordering of topics for the chart that placed most similar topics nearest to each other, where most similar is defined by their distributions over documents. Ideally, these topics should have very similar meanings for the analyst. We then manually reviewed the figure. We noticed that neighboring rows did seem to have similar intensities at similar time periods, validating the overall linear ordering of the topics. We manually placed groups of similar topics into families (1 through 11 in Figure 1). Again, the fact that families of topics did appear contiguously validates the ordering

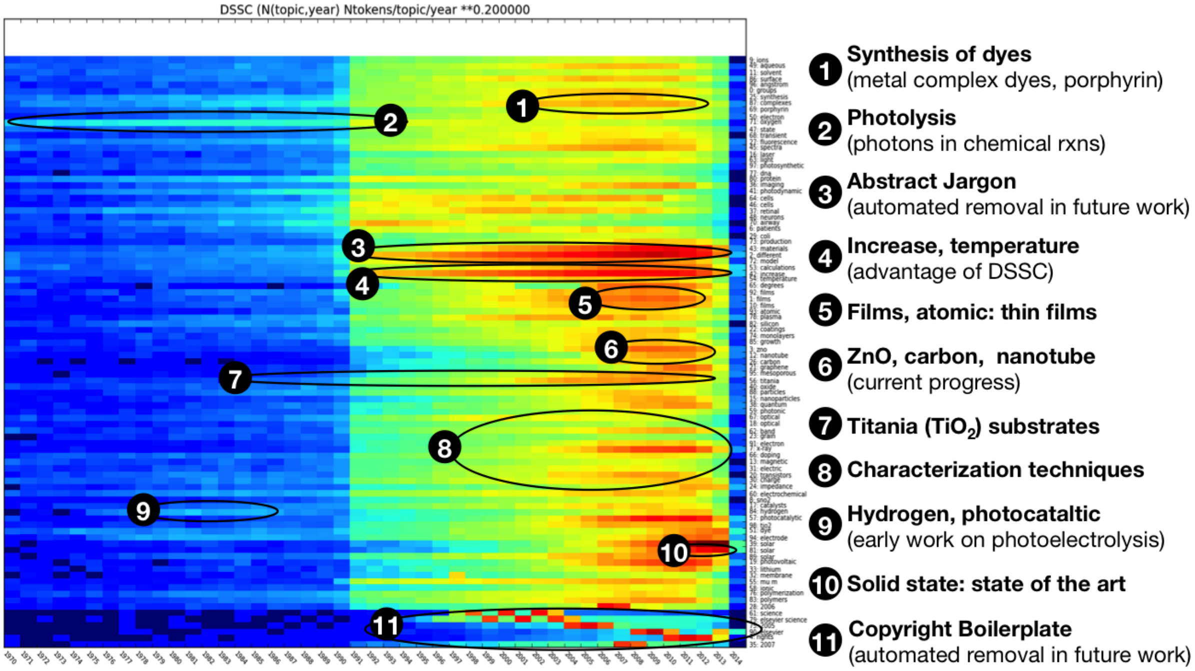


Figure 1: Annotated Topic Model Visualization

An analyst examining Helios output visualized as it is in Figure 1 would have access to the complete term list of each topic (not shown). A domain-knowledgeable analyst should be able to use those full term lists in conjunction with a visualization of topics like that of Figure 1 to identify topic replacement – the shift in research focus from one solution to another.

One of the analysts on the SRI team with a materials science background did this with topics 49 and 81, whose terms lists (for the top 15 terms) were as follows:

Topic 49	**Topic 81**
aqueous	solar
solution	dye-sensitized
solutions	dye-sensitized solar
acid	solar cells
aqueous solution	dye-sensitized solar cells
sodium	dyes
aqueous solutions	cells
surfactant	sensitizers
water	efficiency
acidic	dye
micelles	organic
cationic	dsscs
charged	organic dyes
surfactants	solid-state
sulfate	performance

Table 3: Helios Generated Topic Models

With the benefit of these term lists she recognized Topic 49 as related to liquid electrolytes in DSSCs and Topic 81 as related to solid electrolytes. She then plotted the data for just these two models, as shown in Figure 2. Tokens (Figure 2 y-axis) are units of text that have been isolated from their larger context and identified as distinct units. They usually refer to occurrences of words, though it is possible to tokenize other entities, such as word pairs, words combined with punctuation, etc. Here tokens are occurrences of the terms and phrases associated with each topic of the model. Figure 2 shows a decrease in research related to aqueous solutions while solid-state related research increases. This figure illustrates the ascendance of one technical approach over another and represents a case of potential nascent breakthrough.

The baseline DSSC case study also identified a recent shift towards solid-state-based approaches due to the corrosiveness and other negative features of liquid electrolytes. The topic modeling highlighted the same shift, indicative of its capacity to identify topic replacement. In the work described below for Subtask 4.3 we show how the Helios system can be used to automatically this type of topic replacement.

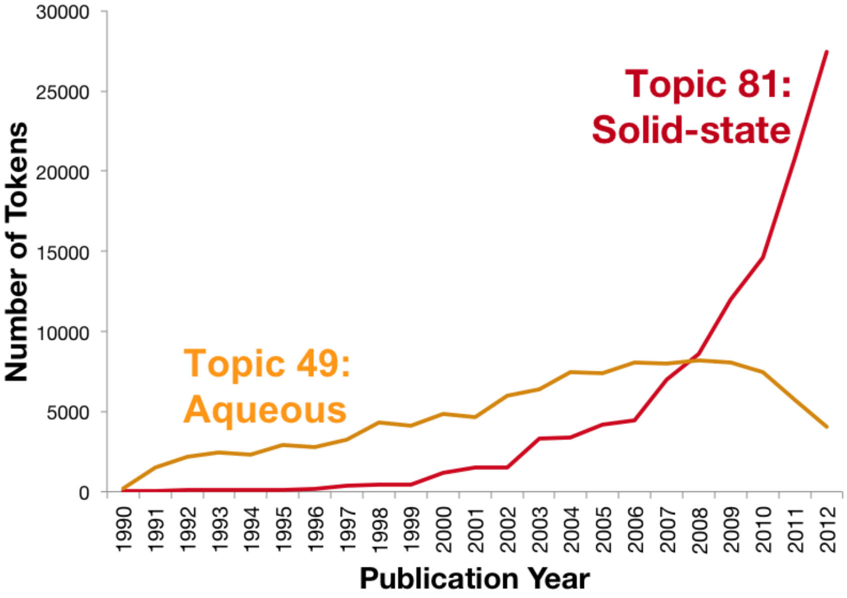


Figure 2: Topic Model Analysis Identifying Shift from Liquid Electrolytes to Solid-State Technology

Table 4 below show the results of the performance assessment; the third column (Target Performance) repeats the pre-project performance targets outlined above in the Budget Period 1 description of the Technical Work Plan section. Table 4 provides the primary quantitative validation of the Helios platform relative to the ground truth constructed via the three human-generated case studies.

DIMENSION	INDICATOR	TARGET PERFORMANCE	ACHIEVED PERFORMANCE	
Coverage	Number of different data sources identified, ingested, and analyzed. (scientific literature databases, patents, technical report databases, etc.)	3-5 data sources	1. Web of Science: 199,821 records ingested and analyzed 2. USPTO: 4,900,938 records ingested 3. NSF Awards: 500,489 records ingested 4. DOE Technical Reports: 8160 records ingested 5. Targeted Full text: 119 records ingested	
Insight	Helios identifies highly influential documents for each technology topic compared against the ground-truth provided by the case studies.	50% Precision 75% Recall, or on par with human	DSSC	52% raw recall 52% corrected recall 2% raw precision 94% ABR-precision
			CdTe	20% raw recall 50% corrected recall 7% raw precision 98% ABR-precision
			MultiJunc	33% raw recall 57% corrected recall 8% raw precision 98% ABR-precision
Insight	Helios recognizes the main solar innovations and their components as illustrated in the case studies.	35% Precision 75% Recall, or on par with human	DSSC	82% raw recall 82% corrected recall
			CdTe	44% raw recall 62% corrected recall
			MultiJunc	46% raw recall 50% corrected recall
Insight	Helios associates the same entities (institutions, authors, etc.) with each innovation as the case studies.	50% Precision 50% Recall	DSSC	84% raw recall 84% corrected recall 100% precision
			CdTe	36% raw recall 57% corrected recall 77% precision
			MultiJunc	50% raw recall 50% corrected recall 100% precision
Efficiency	Helios should be able to process a set of 100,000 documents in less than 24 hours.		Helios can ingest, disambiguate, and analyze approximately 180,000 Web of Science records or 145,000 patent records in 24 hours. Similar levels of performance are achievable on other document types.	

Table 4: Project Data Sets – Document Coverage

Performance results exceeded targets in general for the DSSC dataset, and were roughly equal to targets for the CdTe dataset. We believe that a major reason for this performance differential is that major breakthroughs in CdTe occurred in the 1970s and earlier, where we have very thin coverage of the technical literature in our digital dataset. In DSSCs, where substantial progress has been made in research since 2000, the digital documentary record is more comprehensive and complete.

Raw Recall for documents (row 2 above) indicates system performance compared against all document types and date ranges identified by the human analyst. Corrected Recall indicates system performance compared against the subset of analyst-identified documents from Web of Science published in 1970 or later. The low unadjusted precision rates are driven by false positives. Since virtually any paper in a domain can be important, Precision performance is highly dependent on the population of entities being assessed, and thus the potential for false positives. Because of this phenomenon, the IARPA FUSE program recently replaced all precision program targets with “Adjusted Base Rate Precision” (ABR-precision) program targets. ABR-precision reports what system performance would have been had the original population been 50% targets and 50% non-targets. These adjusted precision values are clearly substantially better. A true measure of precision is probably somewhere between the adjusted and unadjusted number.

Recall scores for innovations (row 3 above) were calculated by mapping the list of major innovations identified by a project analyst (in each case the case study authors) to computer-generated topics. The Raw Recall score indicates system performance matching all innovations identified by the analyst. Web of Science data from the 1970s and before does not include abstract text, making these documents unsuitable for topic modeling. The Corrected Recall indicates system performance matching post 1970 analyst-identified innovations.

The entity analyses in Table 4 (row 4) describes the system’s performance associating entities with each of the specific technologies. The second Insight metric in Table 4 measures how well Helios was able to identify solar innovations. We assessed the topics of our topic model to determine which of these topics were strongly aligned with which solar innovations. It makes sense to ask which innovations are covered by topics; this is the recall number that was reported. Reporting a precision number only makes sense if Helios had a technique for filtering topics in order to determine which corresponded to innovations. The topic model needs to cover all topics in the data, so we never expected it to be restricted to innovations. Since we didn't have filtering in place at the end of Phase 1, no precision score is reported in Table 4 for this entity category. We had originally planned to apply NLP techniques to recognize innovation discussion, but we were unable to implement these in Phase 1. In phase 2 we shifted our focus to looking at topic evolution and especially topic replacement. The work that we did there made progress toward the identification of innovations, but never got far enough to support the type of filtering imagined for the second insight metric.

Subtask 3.4: Entity Tracking

Work on entity tracking was deemphasized as the project progressed. There are no results to report.

Subtask 3.5: Network Influence Analysis

Using Web of Science records, we generated a graph for each technology area depicting the network of paper authors, based on citation, co-citation, and co-authorship. Each graph contained authors as nodes. In the citation graph, a directed edge with weight n is induced from author A to author B if n-many papers by A all cite papers by B. In the co-citation graph, an undirected edge with weight n is induced between author A and author B if A and B are both cited in n different bibliographies. In the coauthor graph, an undirected edge with weight n is induced between A and B if they are coauthors on n papers. We only analyzed the coauthor graphs in this phase.

The coauthor graph for the DSSC document group is displayed in Figure 3. We computed the PageRank of each author in the graph as a means of ranking authors according to their centrality (influence) in the graph. The PageRank of an author is high when an author is cited frequently and when other authors with a high PageRank score cite an author. The PageRank score is assigned to nodes in the network using an iterative algorithm.

We then generated a list of the top 10 authors according to their PageRank score. These lists were then compared to the technology case studies as way to see if Helios could identify the same key authors as those identified by human analysts. For the DSSC case, 8 of the 10 identified by Helios were mentioned in the case study, as depicted in Table 5 below. We would expect similar alignment for the other two technologies (Multi-Junction and CdTe).

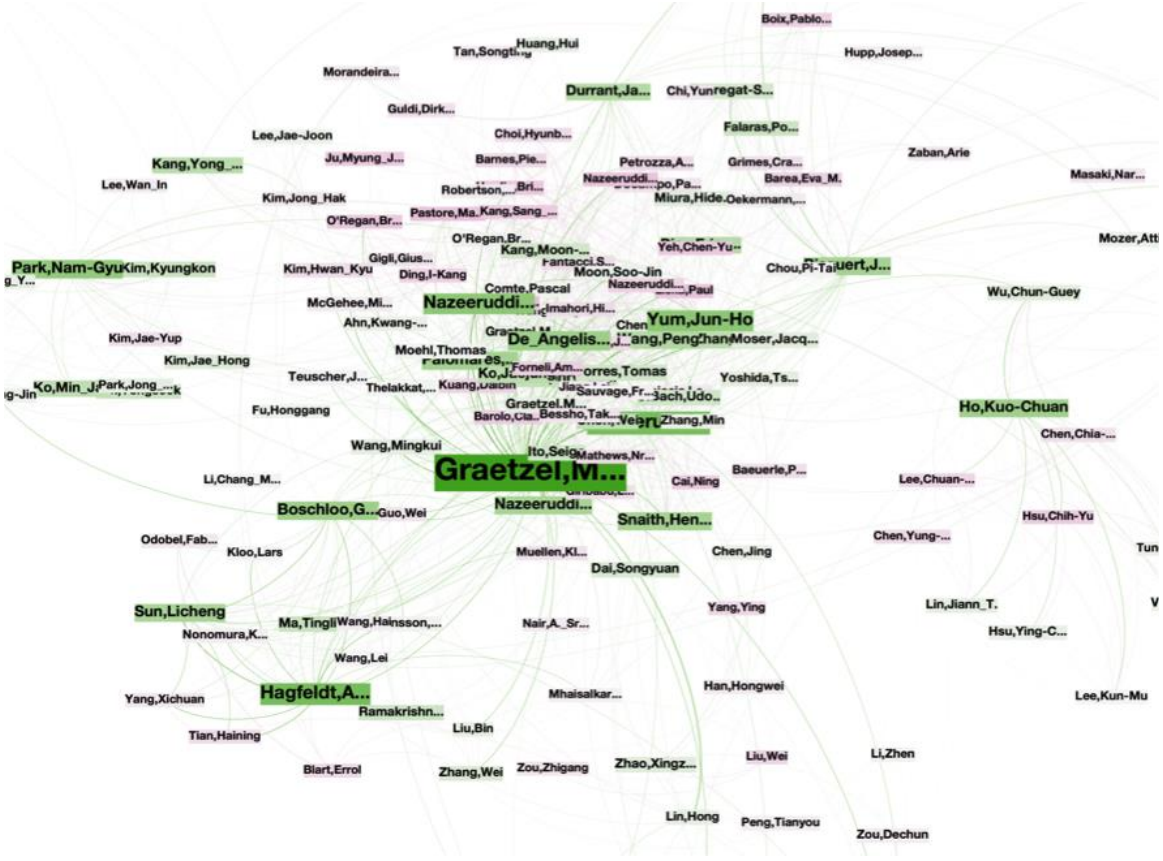


Figure 3: DSSC Research Collaboration Network (1980-2013)

HELIOS Output	*PageRank*	*Included in Case Study*
Graetzel, Michael	*0.001293*	✓
Hagfeldt, Anders	*0.00054*	✓
Zakeeruddin, Shaik_M.	*0.000528*	✓
Yum, Jun-Ho	*0.00043*	✓
Nazeeruddin, Mohammad_K.	*0.000421*	✓
Park, Nam-Gyu	*0.000404*	✓
De Angelis, Filippo	*0.00036*	☒
Boschloo, Gerrit	*0.000355*	✓
Ho, Kuo-Chuan	*0.000349*	✓
Palomares, Emilio	*0.000337*	☒

Table 5: DSSC Research Collaboration Network (1980-2013)

Subtask 3.6: Analysis Team Support

As described in the Introduction Section of this report, the project team consisted of the Technology Analysis Group, drawn from SRI's policy staff and the Advanced Analytics Group, drawn from SRI's Artificial Intelligence Center. These two groups worked closely with one another throughout the project, with the Analytics Team maintaining the Helios platform and running all analyses on the platform while the Technology Analysis Group conducted analyses on outputs from the Helios platform.

Task 4: Helios Platform Extension

Subtask 4.1: Supplemental Data Collection

Following discussions with the SEEDS Project Manager, this task was modified from collecting additional to data to focusing on examining methods for greater cross-dataset integration. The results of text analysis techniques, such as topic modeling, depend significantly on the literature being analyzed. Different types of technical documents (e.g., scholarly articles, patents, technical reports, grant abstracts) will use different terminology and phrasing to describe the same phenomena. In each type of document, the authors are likely to adopt terms and conventions appropriate for the audience likely to read the document, and aligned with the standards of rhetoric and composition for the profession using the document. Consider the difference between scientific scholarly articles and patents.

In a scholarly article on a given technology, the primary authors are researchers who are describing their own work and any resulting discoveries. They are communicating with peer experts within their discipline (most commonly), and therefore use terms specific to that discipline, without including much in the way of explanation for those terms. The motivation of the author is to make clear what is significant about the discovery so that others in the field will recognize the value of that contribution to the literature.

For a patent regarding that same technology, the primary authors are patent agents. Most of these are writing the patent based on a draft description by the inventor—in some significant but rare cases, the inventor writes the patent claims. In any event, the patent application is being written for review by a patent examiner, who is likely to be familiar with the technology but not necessarily steeped in its technical details. The claims are worded to address the specific criteria used to evaluate patents (novelty, non-obviousness, utility), but also to obfuscate the true significance of the invention for competitive reasons.

A second confounding factor in tracking topics across these document sets are the differences in the subject matter and timing of topics across patents and papers. Patents describe inventions that have been “reduced to practice,” while papers describe laboratory discoveries. Therefore, papers are likely to including topics related to more basic science and the research process, while patents are more likely to focus on tangible concepts that can be implemented as technologies. As a result, not only will terminology differ, but the proportion of attention placed on different types of concepts will differ as well. Also, in patenting, applicants tend to favor explaining new concepts in great detail until they become established knowledge. In papers, researchers have a tendency to invent new vocabulary as a means of creating their own “niche” in the community of practice around that concept.

The two established methods of linking the content of article and patent datasets are (1) using non-patent literature (NPL) citations, where a scientific article is listed as a reference in the patent document, and (2) creating a disambiguated author-inventor database, where researchers who author articles can be linked to any patents on which they are listed as an inventor, and vice versa. Both methods have limitations.

For NPL citation, the first barrier is that patents do not follow a particular citation convention (e.g., APA or IEEE) when listing references. Therefore, researchers have to identify any reference that is a paper, extract it, convert it into a standardized format, and then see if it is present in an article database (e.g, Scopus or WoS). This can be very laborious, so some NPL instances may be missed. Also, references to NPL are highly domain dependent. In some fields, like pharmaceuticals, patents often cite NPL articles, due to the strong pipeline from basic research to drug development. In other domains, such as automotive engineering, articles appear rarely as NPL.

For author-inventor linkages, the quality of the disambiguation and association between names in the USPTO and article dataset determines the quality of the linkage. A high-quality database has been created by a research team led by the Harvard Institute for Quantitative Social Science, linking USPTO records with the MEDLINE database. This is limited primarily to biomedical research literature. Variations in the spelling of names, including the inclusion or exclusion of middle names, could confound the linkage. Also, the linkage only works in domains where researchers are likely to be co-inventors on patented technology. In domains where the path from research to commercialization is less direct, this approach has little use.

With these challenges in mind, we attempted to develop a method for aligning patents with papers in our data. We started by using the search queries developed for the Web of Science dataset to partition the other four datasets into document groups for each of the three target technologies. The document counts for each group are shown below in Table 6. Note that for most document groups, the counts tend to be larger in more recent years. This is because there is greater availability of digitized document records in recent years and because the literature for a given technology tends to expand as the technology matures.

Source	Document Group Sizes[i]	
Thomson Scientific Web of Science	MJ:	4104 (32362)
	CdTe:	4238 (38469)
	DSSC:	15046 (139270)
U.S. Patent & Trademark Office	MJ:	622
	CdTe:	220
	DSSC:	942
U.S. Dept. of Energy	MJ:	286
	CdTe:	420
	DSSC:	50
U.S. National Science Foundation	MJ:	97
	CdTe:	43
	DSSC:	133
Thomson Scientific & others	MJ:	28
	CdTe:	14
	DSSC:	77

Table 6: Partitioned Dataset Statistics

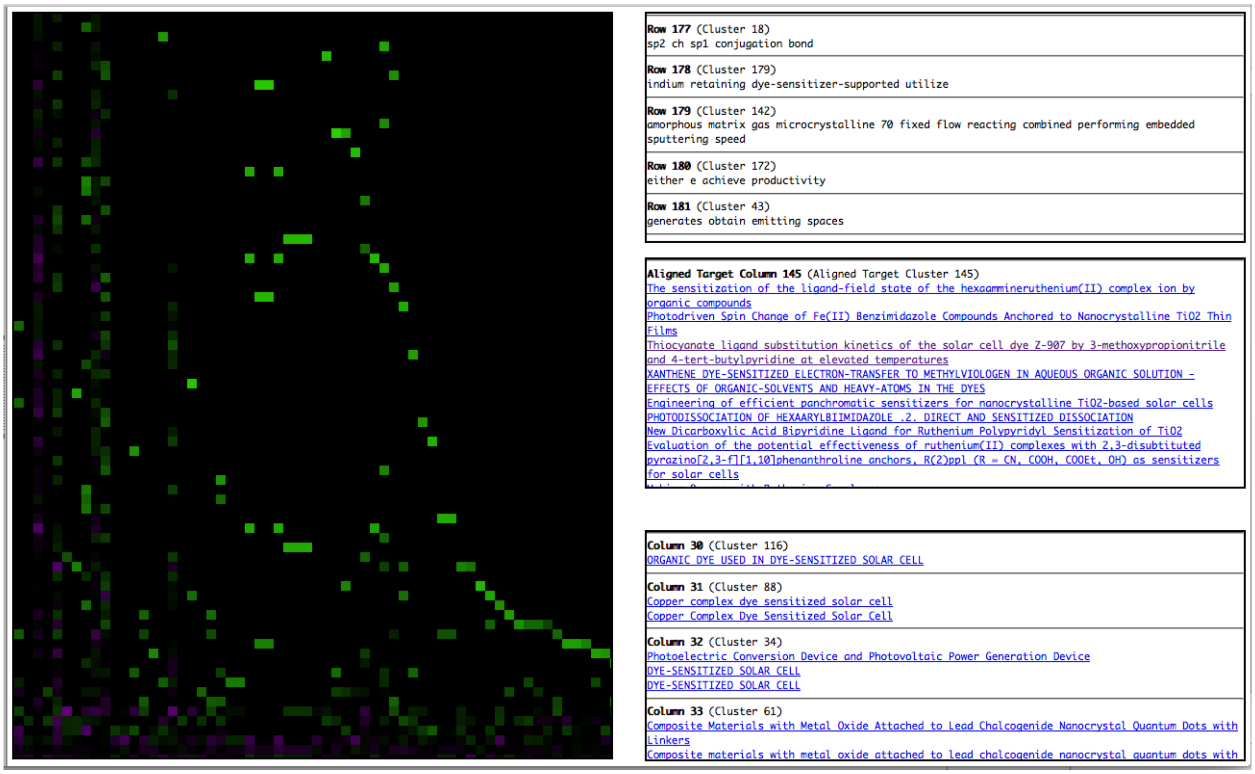
Next we attempt to reconcile patents to scientific papers in the following process:

1. We take the USPTO (patent) dataset for a given technology, and derive 200 “term clusters,” and 200 associated “patent clusters.” The patent clusters are sets of patents that are topically related, based our co-clustering analysis. The term clusters are the specific words that distinguish among the patents in each cluster—in other words, the terms that are most unique to the topic associated with that cluster. Note that a very large share of patent clusters were one-patent clusters.

[i] For Web of Science, first number is the document record count. The number in parentheses is the count of citation “stubs” (works cited by papers in the document group).

2. Using the term clusters derived in Step 1, we “compress” the Web of Science (paper) dataset to assign WoS documents to each of the 200 term clusters, based on the appearance of the terms in a cluster to particular WoS papers.
3. Using those same term clusters, we can now associate clusters of WoS papers to each of the 200 patent clusters. Note that in doing so, the same patent cluster can be related to multiple paper clusters.
4. We then display these relationships in a grid, where each row is a term cluster and each column is a patent cluster. Where the patent cluster is most strongly associated with a term cluster, that cell in the grid is highlighted; purple shading indicates negative correlation.

In essence, we attempted to align groups of related patents area to groups of related papers based on the similarity of their terms. Selecting a cell will display the corresponding term cluster, patent cluster, and paper cluster. Figure 4 below presents a screenshot of the visual display for the patent-to-paper topic alignment. The display was linked to the Document Browser, so that an analyst can observe the terms in the cluster, and manually inspect the associated patents and papers to evaluate their correlation. Green indications high correlation between patent and term clusters; purple indicates negative correlation, e.g. terms in a given cluster were statistically less likely to appear in a given patent cluster than across all clusters on average.



Row 177 (Cluster 18)
sp2 ch sp1 conjugation bond

Row 178 (Cluster 179)
indium retaining dye-sensitizer-supported utilize

Row 179 (Cluster 142)
amorphous matrix gas microcrystalline 70 fixed flow reacting combined performing embedded sputtering speed

Row 180 (Cluster 172)
either e achieve productivity

Row 181 (Cluster 43)
generates obtain emitting spaces

Aligned Target Column 145 (Aligned Target Cluster 145)
The sensitization of the ligand-field state of the hexaammineruthenium(II) complex ion by organic compounds
Photodriven Spin Change of Fe(II) Benzimidazole Compounds Anchored to Nanocrystalline TiO2 Thin Films
Thiocyanate ligand substitution kinetics of the solar cell dye Z-907 by 3-methoxypropionitrile and 4-tert-butylpyridine at elevated temperatures
XANTHENE DYE-SENSITIZED ELECTRON-TRANSFER TO METHYLVIOLOGEN IN AQUEOUS ORGANIC SOLUTION - EFFECTS OF ORGANIC-SOLVENTS AND HEAVY-ATOMS IN THE DYES
Engineering of efficient panchromatic sensitizers for nanocrystalline TiO2-based solar cells
PHOTODISSOCIATION OF HEXAARYLBIIMIDAZOLE .2. DIRECT AND SENSITIZED DISSOCIATION
New Dicarboxylic Acid Bipyridine Ligand for Ruthenium Polypyridyl Sensitization of TiO2
Evaluation of the potential effectiveness of ruthenium(II) complexes with 2,3-disubtituted pyrazino[2,3-f][1,10]phenanthroline anchors, R(2)ppl (R = CN, COOH, COOEt, OH) as sensitizers for solar cells

Column 30 (Cluster 116)
ORGANIC DYE USED IN DYE-SENSITIZED SOLAR CELL

Column 31 (Cluster 88)
Copper complex dye sensitized solar cell
Copper Complex Dye Sensitized Solar Cell

Column 32 (Cluster 34)
Photoelectric Conversion Device and Photovoltaic Power Generation Device
DYE-SENSITIZED SOLAR CELL
DYE-SENSITIZED SOLAR CELL

Column 33 (Cluster 61)
Composite Materials with Metal Oxide Attached to Lead Chalcogenide Nanocrystal Quantum Dots with Linkers
Composite materials with metal oxide attached to lead chalcogenide nanocrystal quantum dots with

Figure 4: Aligning Patents and Papers via Term Co-Clustering

Assessing the degree of alignment was done manually by selecting highlighted cells and evaluating whether the patent and paper clusters associated with that term cluster were indeed similar. We examined 20 cells chosen at random. For each cell, we examined the first five papers identified as matching that term cluster for similarity to the first patent associated with the term cluster; as indicated above, there was often only one patent in the patent cluster.

Figure 5 below shows the distribution of alignment achieved, where the y-axis is the portion of papers that seemed to match the topic area of the associated patent, with a range of zero to fine (0% to 100%). The horizontal labels of Figure 5b report the row and column number of the cell examined (where the origin cell is 0,0) and the associated paper cluster number. So the leftmost bar in Figure 5 looked at the cell 1,1 and the document cluster that was matched to that cell, cluster 145. In this case, four of the first five papers seemed to match the single patent in the associated patent cluster.

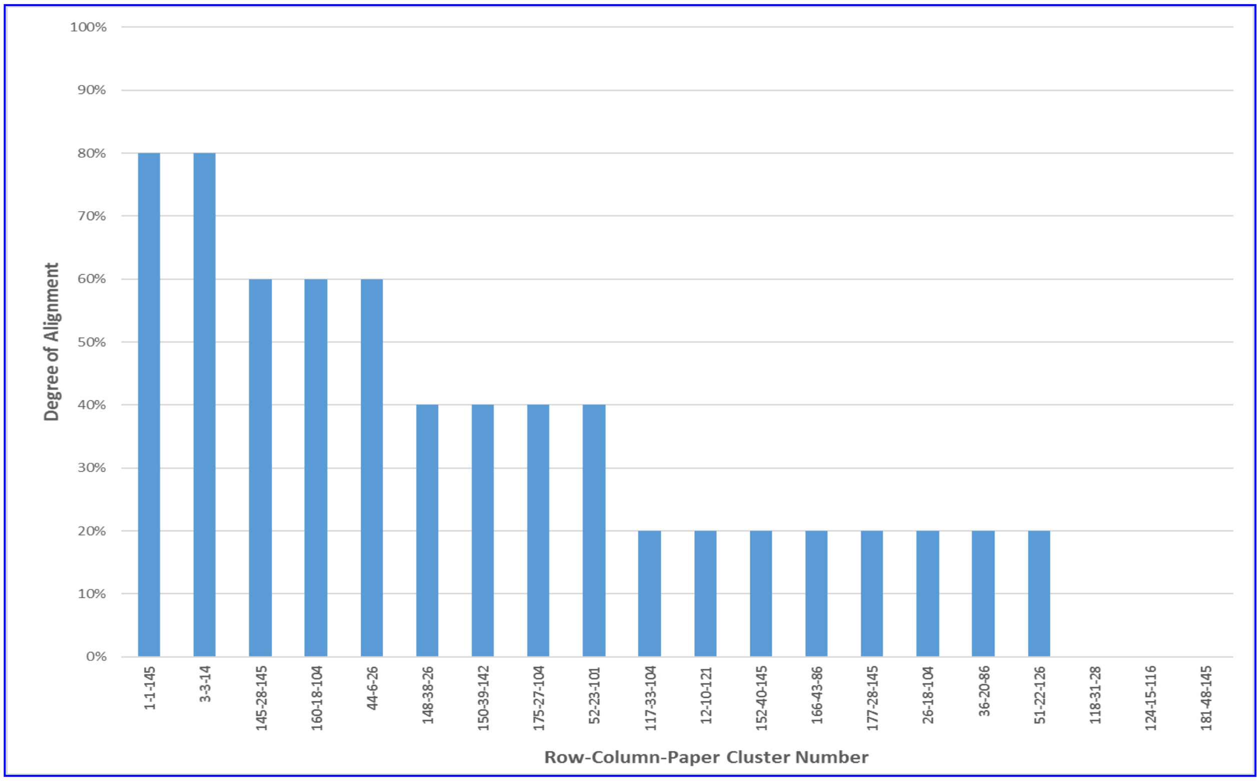


Figure 5: Patent-Paper Alignment Results

As expected, overall there are cases of fairly strong alignment and also evident misalignment in the outputs from this experiment. Given the value of being able to reliably align across disparate document groups, it would be useful to develop a method to automate alignment assessment. The project team explored this but was did not make any reportable progress. It remains an item for future work.

The very partial assessment that was done indicates that the alignment could be better. The current technique is only the first equation tried, which is one that uses all groups of terms but weights them by their pointwise mutual information (PMI) with document clusters. An improvement to this algorithm would likely be to look only at the most important term clusters and drop the others completely for making decisions about topic alignment. Under this approach it would likely make sense to look at statistical significance as well as PMI in order to rule out spurious correlations.

Subtask 4.2: Data Ingest and Disambiguation

The team at SRI Advanced Analytics completed Extract, Transform & Load (ETL) runs on all five datasets so that they are contained fully in the Helios database. The team also implemented an application to enable index-based searching of each dataset.

Subtask 4.3: Advanced Topic Modeling

The basic topic modeling of Period 1 was expanded in to include consideration of how Helios could be used to examine topic evolution over time and how the human-identified topic replacement of Subtask 3.3 could be identified by machine. We also examined ways to automate the assignment of meaningful topic model names using Wikipedia category labels.

Topic Evolution

The project's initial topic modeling described above for Subtask 3.3 analyzed entire document sets and then looked over their entire time horizon for indications of technological progress. This approach provides for reasonable retrospective study, but it shapes a vocabulary of terms based on all past usage up through the most recent publication date. We know that scientific terminology and concepts change over time, as new terms are introduced or reconfigured to describe new phenomena (such as a new analytical technique). As a result, a vocabulary learned over documents from a specific earlier time period will differ from a vocabulary learned from documents in a later time period.

For example, analysis of documents from 1980 through 2010 will show certain patterns of topics occurring in documents published in 1990. Analysis of documents from 1980 through 2000 would include all of the same 1990 terms and documents, but they would be arranged differently (because they lack influence from the 2001-2010 documents). Therefore, the terms and term clusters themselves learned using the 1980 to 2000 document set would look very different from those generated using the 1980 to 2010 document set. As a result, our initial topic modeling does not fully convey the temporal context of documents, as documents from any given year are reflective of the state of the scientific terminology as it existed at that time, while our analysis takes into account terminology and usage that emerged in later publications.

In order to gain a better understanding on how topics evolve over time, we generated a sequence of clusters of Web of Science abstracts, where each cluster covers a window in time. This approach provides us with a series of chronologically ordered "snapshots" of scientific terminology, similar to frames in a film reel. Comparing results across these windows allows us to limit the generation of term clusters to those that were in contemporaneous use at the time of the most recent publication in each window. We experimented with different time periods and degrees over overlap between adjacent time periods. Qualitatively, a set of six-year periods with three years of overlap provided the best initial results.

We then look at the movement of terms from one cluster (window) to the next. Overall, because of the significant overlap in documents from one window to the next, we expect adjacent clusters to be highly consistent. However, we expect to see some clusters gradually disperse and new clusters gradually form over time. The use of successive overlapping windows enables our technique to highlight those patterns that are “new” as we move forward in time from the earliest window, thus isolating the degree of change in topics from one year to the next.

Each column/color in Figure 6a below represents a topic model built on a six-year time slice of the DSSC publication data set (the figure is truncated above and below given the range of data). A pair-wise similarity metric is computed between all topics time slice and the subsequent time slice, and is shown in a binary fashion as a binary line connecting topics between time periods.

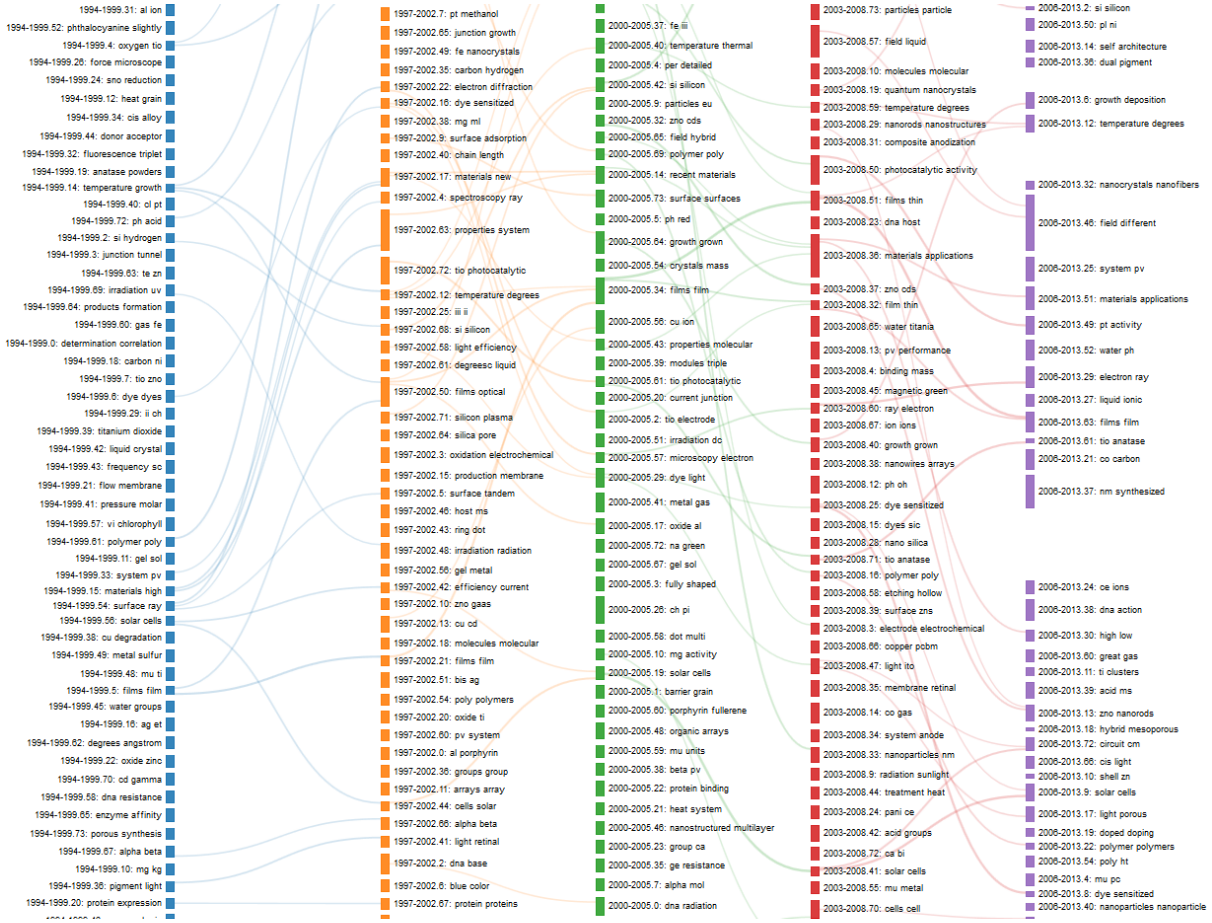


Figure 6a: 6-Year Overlapping DSSC Topic Models

(Color coding only meant for ease of visualizing separate 5-year periods)

By clicking on a single topic in a given time slice, an analyst can identify related topics in previous and subsequent time slices. In Figure 6b below, a hypothetical analyst has selected a topic related to “dyes” in the 2000-2005 time slice. Unlike Figure 6a in which topic similarity is shows in a binary fashion with a line drawn between any two topics that have some threshold level of similarity, here topic similarity is represented as a shaded line with increasing thickness corresponding to greater similarity. Once can see that the dyes topic is closely related to the “dye sensitized” topic in the subsequent time slice.

More generally, Figure 6b shows how the time period specific topic modeling can be used to show how different topics coalesce into single topics over time, and how single topics can diverge, though there are limitations to this approach.

The three-years of overlap between time periods was expected to stabilize the models and allow for meaningful comparisons. While this did occur to some extent, we also saw a much larger amount of topic splitting and merging than we had expected. We attribute this not to actual changes in the data but to the instability of the topic modeling algorithm. Topic modeling uses a stochastic search algorithm, and small changes in the input data set can lead to very different final results. We would like to generate more stable algorithms in the future so that small changes in the input only lead to small changes in the output. One way to accomplish this might be to remove all cluster assignments that are not statistically significant. This could lead to a stable core. The terms that had been removed could then be re-inserted greedily rather than stochastically.

Another approach to obtaining more stable model without a more stable base algorithm is to start with a fixed topic model for a given reference period, and then to extend the model forward and backward in time by making initial assignments based on the established model and then optimizing greedily to obtain the model for the new period. Both of these approaches remain to be done under future work.

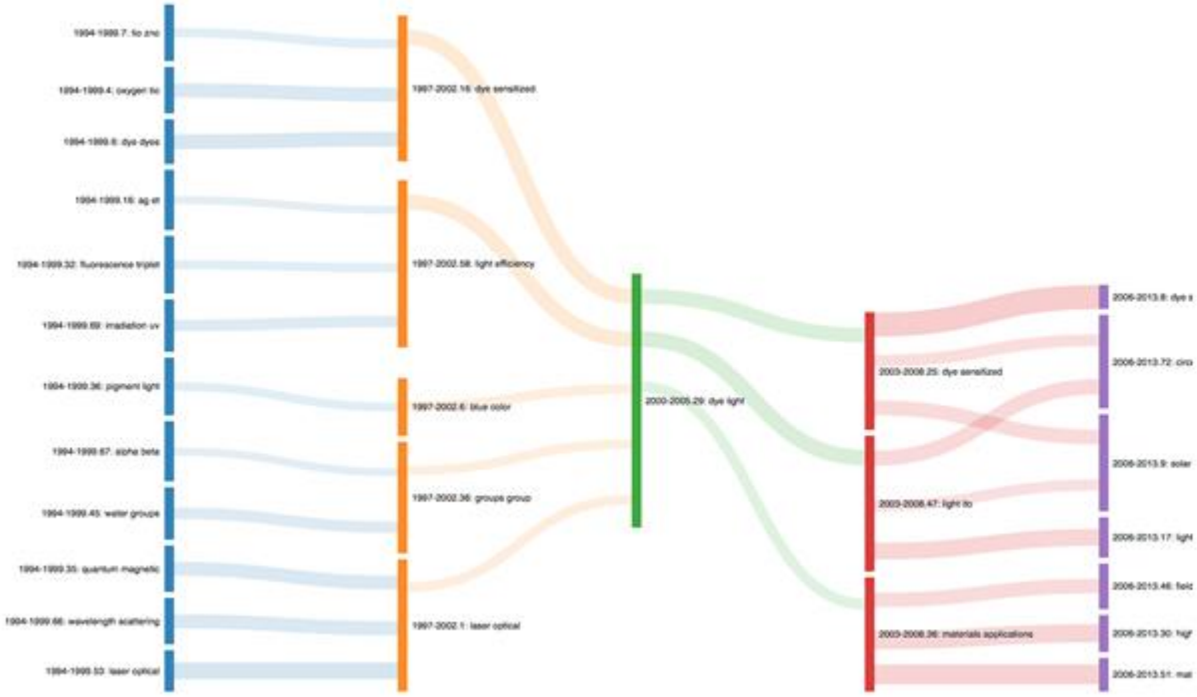

1994-1999.7: tio zno
1994-1999.4: oxygen tio
1994-1999.8: dye dyes
1994-1999.16: ag el
1994-1999.32: fluorescence triplet
1994-1999.69: irradiation uv
1994-1999.36: pigment light
1994-1999.67: alpha beta
1994-1999.45: water groups
1994-1999.35: quantum magnetic
1994-1999.66: wavelength scattering
1994-1999.53: laser optical
1997-2002.16: dye sensitized
1997-2002.58: light efficiency
1997-2002.6: blue color
1997-2002.36: groups group
1997-2002.1: laser optical
2000-2005.29: dye light
2003-2008.25: dye sensitized
2003-2008.47: light ito
2003-2008.36: materials applications
2006-2013.8: dye s
2006-2013.72: circ
2006-2013.9: solar
2006-2013.17: light
2006-2013.46: field
2006-2013.30: high
2006-2013.51: mat

Figure 6b: 6-Year Overlapping DSSC Topic Models

(Color coding only meant for ease of visualizing separate 5-year periods)

Automated Topic Replacement Identification

In the Figure 6 equivalent for CdTe, the topic modeling seemed to show neighboring families (i.e., similar groups of topics) in which one family was focused earlier in time and the other was focused later. In some cases, the earlier family appeared to be related primarily to study of natural phenomena (more basic research) and the later family related to the application of the earlier phenomena. For example, topics on films and substrates followed (and were adjacent to) topics on epitaxy and doping. Similarly, topics on quantum dot phenomena were followed (and were adjacent to) topics on quantum dot manufacturing.

This observation suggests that we should (a) see asymmetric patterns of citation between such topics, and (b) be able to use this asymmetry to automate identification of topic replacement. Figure 7a below show illustrates the citation patterns of documents associated with 100 topics within DSSC (the number 100 here is arbitrary). The same 100 topics are listed vertically and horizontally, and the color at any point on the plot measures the degree to which documents in one topic cite documents in the other. Average citation rates are indicated by light green shading. Yellow to orange to red indicate stronger correlation whereas darker green to blue indicate declining correlation.

Figure 7a shows several of instances of asymmetric correlation, a number of which are circled. The pair circled in the top right corner are Topic 82 and Topic 78. Table 7 below shows some of the key terms in these two topics, which we have manually labeled

Amorphous Silicon PV (82) and Nanocrystaline Silicon PV (Topic 78). Notice in Figure 7a that these two topics are adjacent to one another, meaning they are similar to one another (see discussion of topic modeling in Subtask 3.3 above). Notice too that within this pair, the documents in the nanocrystaline silicon Topic (Topic 78) frequently site the papers in the amorphous silicon Topic (Topic 82), whereas papers in the amorphous silicon topic cite those of nanocrystaline silicon quite rarely.

When we look at these two topics over time in Figure 7b below, we that Topic 82 was active early, but fades out as 78 becomes active. In other words, over time, nanocrystaline approaches in PV superseded amorphous silicon approaches, which is what happened within in the actual PV research world.

Topic 82 Amorphous Silicon	**Topic 78** Nanocrystaline Silicon
silicon	plasma
amorphous	microcrystalline
porous silicon	amorphous
passivation	plasma-enhanced chemical vapor
hydrogenated	nanocrystalline silicon
annealing	dilution

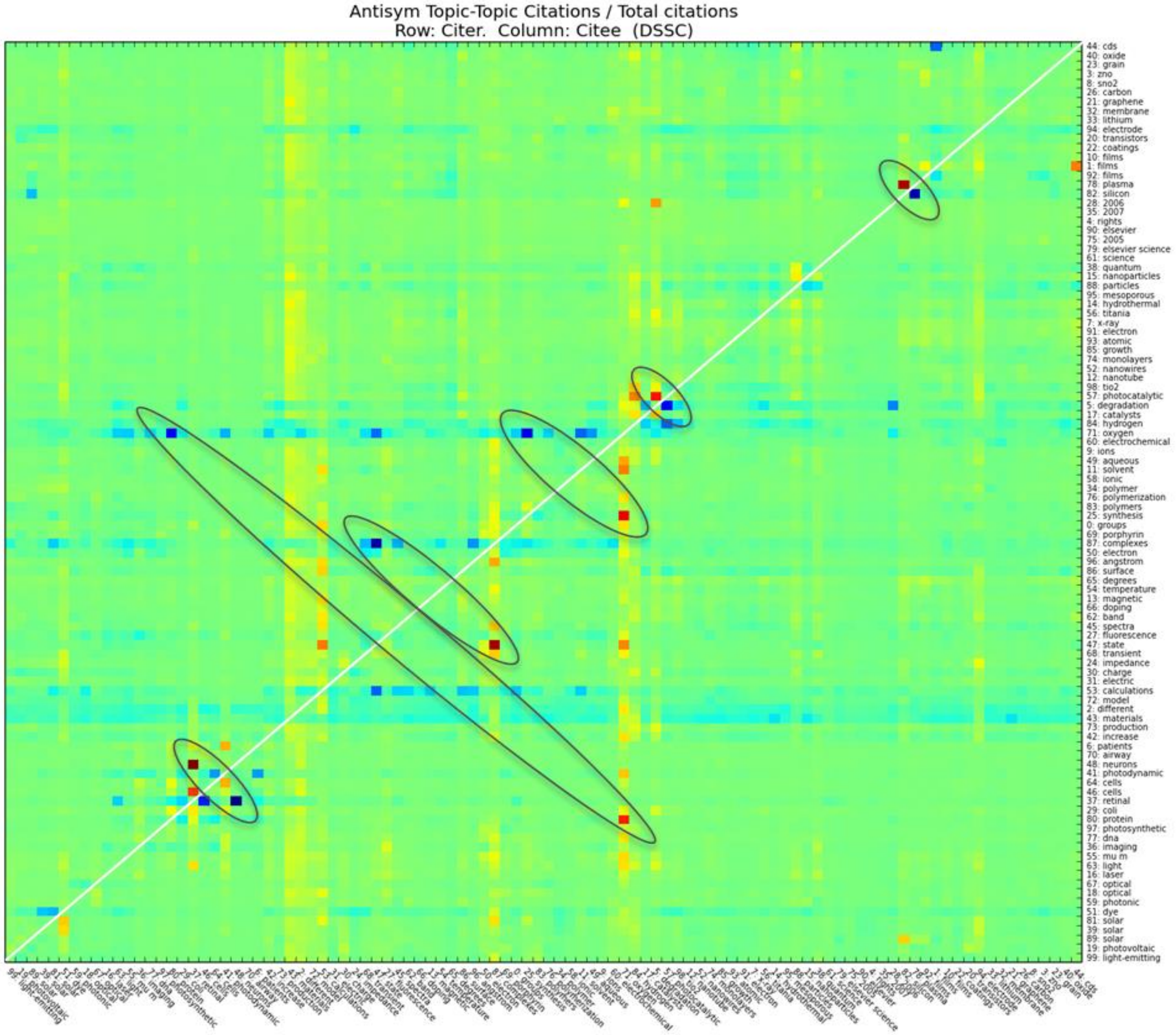
Table 7: DSSC Research Collaboration Network (1980-2013)

This example is not completely automated. Even though Helios identified the unusual asymmetric pattern of citation between these two related topics, a human analyst familiar with the technology was involved in confirming that this was a case of topic replacement. In principle, a fully automated process could be created if it could show the following:

1. That Topic A and Topic B both refer to the same phenomenon (e.g., a method of cell fabrication)
2. That Topic B appears later in the literature than Topic A
3. That Topic B appears in a substantial number of works citing Topic A, but works on Topic A rarely cite Topic B

4. That Topic B is associated with a claim of superior technical performance of some kind relative to Topic A

On point 4 above, we did significant work in the area of performance claims extraction, discussed below for Subtask 4.5, Additional Indicators.



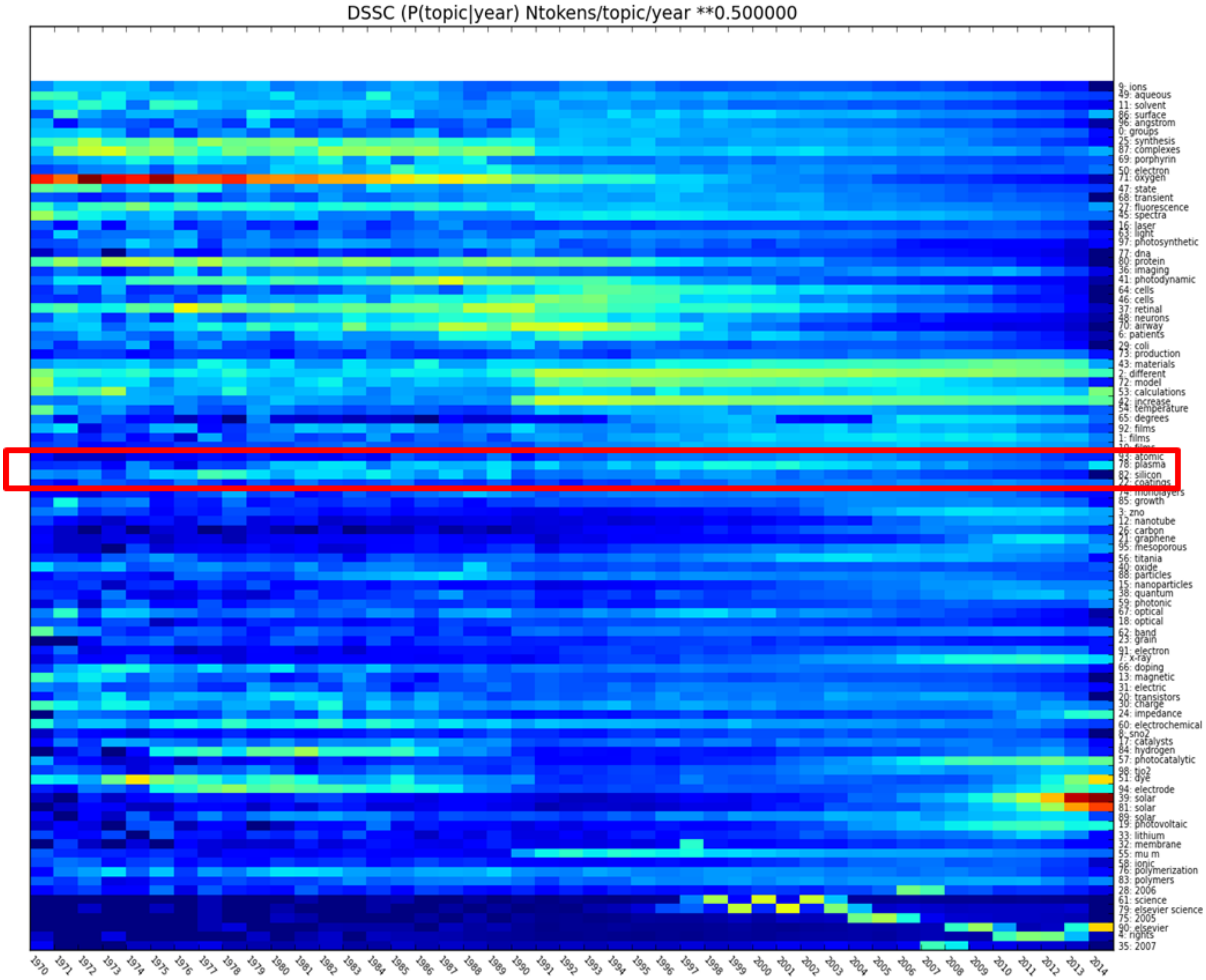


Figure 7b: DSSC Topic Activity over Time

(Shading indicates the number of occurrences of that topic in that year, ranging from dark blue {few} to red {many})

Automated Topic Naming

One limitation of our work on topic modeling, and indeed current techniques for clustering large sets of documents by machine-generated topics, is that the topics in each model are expressed as a simple string of terms. The terms in a given topic are those generated by the model as being most descriptive of the content of the documents in that cluster. These raw term strings are not placed in any particular context, and so they appear to an uninformed observer as a random collection of words. Even a domain-knowledgeable analyst may need to scrutinize a set of terms closely to deduce its relation to a scientific topic a human would understand as a topic, and then he or she needs to give that topic a name.

In an attempt to make the outputs of the topic modeling algorithms more useful to analysts and especially to non-technical managers, we developed a method for finding contextually meaningful topic names based on Wikipedia category labels. For a given topic derived from the WOS documents, we match its terms with one or more Wikipedia articles that are most similar in their subject matter.

Specifically, we take the documents associated with a given topic, extract the top terms from that topic, and then search the documents for sentences where each term appears. We then extract (roughly) the entire sentence where each term appears, and then concatenate those sentences to form a "pseudo-document." This "document" is essentially a string of sentences containing the topic terms. We compare that pseudo-document to Wikipedia articles using Latent Semantic Analysis. We can then calculate the set of Wikipedia articles most similar to the terms in the topic being analyzed. The initial analysis is shown below in Table 8.

Topic ID	Topic Terms	Wiki Article Titles	Wiki Categories
1994-1999.56	solar cells cell efficiency current silicon voltage si cm photovoltaic	Theory of solar cells Solar cell Protocrystalline	Solar cells Thin-film cells Energy conversion
2006-2013.16	dyes operation meso flow photocatalysis color bodipy simulator et red	CZTS Dye-sensitized solar cell Perovskite solar cell	Solar cells Photovoltaics Thin-film cells
2006-2013.60	great gas based sensing sensor sensitivity high sensitive dc sensors	Residual gas analyzer Inductively coupled plasma mass spectrometry Induction plasma technology	Mass spectrometry Laboratory techniques Gas chromatography
2000-2005.17	oxide al ito tin sputtering indium transparent glass coating oxides	Copper indium gallium selenide solar cells Indium tin oxide Transparent conducting film	Solar cells Thin-film cells Silicon forms

Table 8: Using Matched Wikipedia Topic Labels for Modeled Topic Labels

The initial results shown above are promising. Upon expansion of this experiment, however, we found that irrelevant articles began to be associated with the topics generated by the Web of Science articles. It appears that since terms are used in different contexts, and it is not easy for the machine algorithm to understand when a term is being used in reference to solar photovoltaic technology, or when it is used in an entirely different subject domain. Therefore, we found a number of topics were being labeled with Wikipedia-based tags that were clearly not related to solar PV.

In a future work, instead of using the entirety of Wikipedia to match articles to topics, one could use only the subject categories in Wikipedia related to science, technology and energy, or more generally, one could filter the Wikipedia categories to those most relevant to the domain being modeled and use only articles from those categories for the matching.

One could also potentially use a method of Association-Grounded Semantics to label topics more precisely. In this approach, one would start with an existing taxonomy or thesaurus of terms that is descriptive of general topics (e.g., “thin-film”) and then find an external reference that provides a few additional terms that are associated with that term. Those additional terms comprise the “language model” for that term—a more detailed description of the meaning of that term. One could then run a topic co-clustering algorithm on the Web of Science documents to expand each language model, and then cluster the documents by the degree of similarity of the documents to each term’s language model. This approach would require fairly intensive use of human and computational resources.

Subtask 4.4: Advanced Entity Tracking

Work on entity tracking was deemphasized as the project progressed. There are no results to report.

Subtask 4.5: Additional Indicators

We used Helios to explore two specific sets of expanded indicators, one that measures technical advance via technological claims analysis, and another that examines the role of geography of PV research.

<u>Technological Achievement Claims Analysis</u>

In the course of developing the human-generated case histories, the Technology Analysis Team noted that a useful method for detecting and tracking apparent breakthroughs in PV technology is to look at the specific claims of improved technical performance that a research team makes when it reports a particular new development. These claims are often highlighted in descriptions of breakthroughs appearing in the non-technical literature. Take, for example, a recent press release from the Massachusetts Institute of Technology, summarizing the results of a research project on a new technique for constructing solar cells:

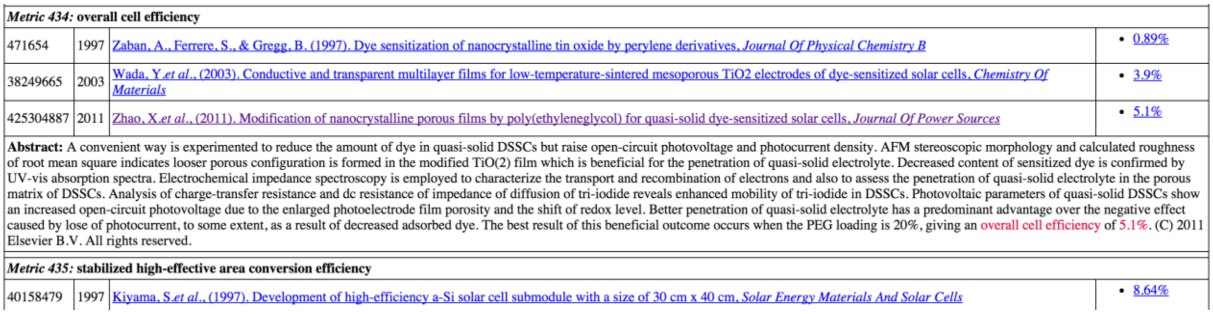
> Researchers at MIT and Stanford University have developed a new kind of solar cell that combines two different layers of sunlight-absorbing material in order to harvest a broader range of the sun’s energy. The development could lead to photovoltaic cells that are more efficient than those currently used in solar-power installations, the researchers say… In this initial version, the efficiency is 13.7 percent, but the researchers say they have identified low-cost ways of improving this to about 30 percent — a substantial improvement

> over today’s commercial silicon-based solar cells — and they say this technology could ultimately achieve a power efficiency of more than 35 percent. (Chandler, 2015)

We developed an approach to enable the construction of “technical claims” charts similar to the NREL “Best Research Cell Efficiencies” chart (http://www.nrel.gov/ncpv/images/efficiency_chart.jpg). A member of the Technology Analysis Team reviewed several full-text technical scholarly articles and identified examples of paragraphs or sentences where authors made claims that their particular solar cell variant achieved a superior level of performance relative to other variants, or relative to the maximum achieved performance reported previously. Humans can spot such clauses readily, as they generally conform to a syntactic pattern such as, “We fabricated a [type of solar cell] that yields a [performance parameter] of [value][unit].”

To develop a semi-automated process for extracting examples of performance metrics from documents, the Platform Development Team deployed the BRAT[ii] tool and embedded coding that allowed the Technology Analysis Team to isolate performance claims in a given full-text document, and then label the phrases in that claim corresponding to the coding structure previously developed. This involved configuring the tool to identify patterns similar to those in the hand annotated examples from the FULLTEXT dataset. The SEEDS Web of Science dataset was used for testing. When the tool is run it generates its output into a structured text file.

The Platform Development Team leveraged work performed previously for IARPA under the ForeST (Forecasting Science and Technology) project, where performance metrics were often used to construct predictions of technical advances that could then be evaluated by participants in an online prediction market. For Helios, the structured text file was used as the input for a HTML report generator, which constructs a single HTML table with sub-sections for each identified metric, see Figure 6 below. The sub-sections list the documents in which the metric was found as well as the corresponding values extracted (right hand column of Figure 8a).

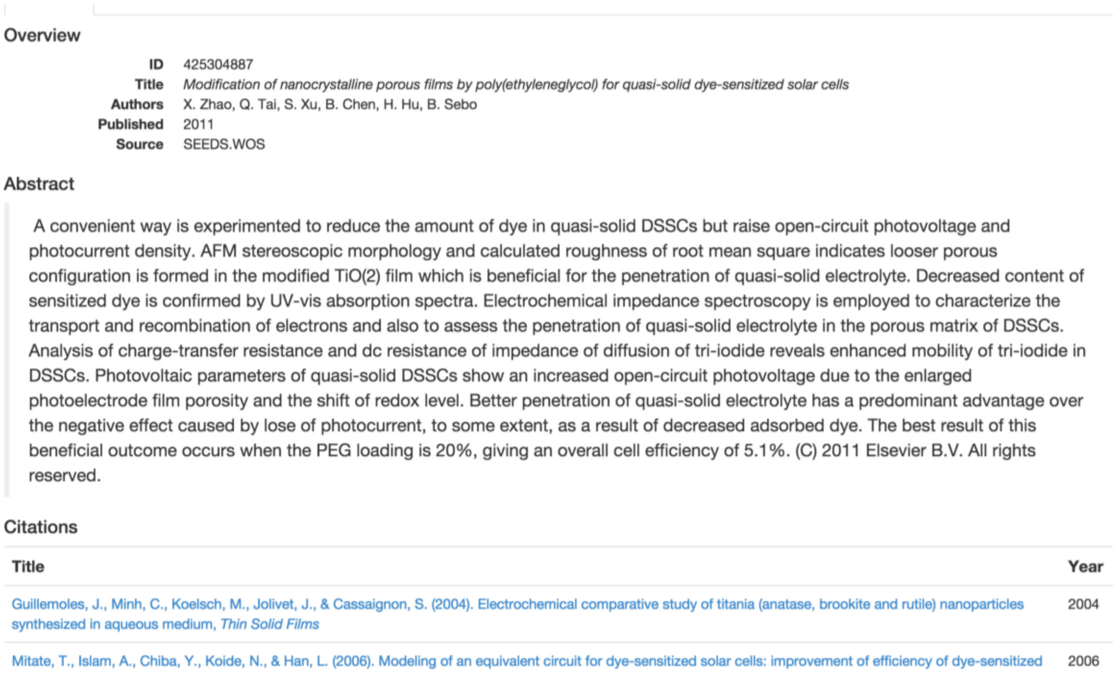


***Metric 434:* overall cell efficiency**			
471654	1997	Zaban, A., Ferrere, S., & Gregg, B. (1997). Dye sensitization of nanocrystalline tin oxide by perylene derivatives, *Journal Of Physical Chemistry B*	• 0.89%
38249665	2003	Wada, Y.*et al*., (2003). Conductive and transparent multilayer films for low-temperature-sintered mesoporous TiO2 electrodes of dye-sensitized solar cells, *Chemistry Of Materials*	• 3.9%
425304887	2011	Zhao, X.*et al*., (2011). Modification of nanocrystalline porous films by poly(ethyleneglycol) for quasi-solid dye-sensitized solar cells, *Journal Of Power Sources*	• 5.1%
Abstract: A convenient way is experimented to reduce the amount of dye in quasi-solid DSSCs but raise open-circuit photovoltage and photocurrent density. AFM stereoscopic morphology and calculated roughness of root mean square indicates looser porous configuration is formed in the modified TiO(2) film which is beneficial for the penetration of quasi-solid electrolyte. Decreased content of sensitized dye is confirmed by UV-vis absorption spectra. Electrochemical impedance spectroscopy is employed to characterize the transport and recombination of electrons and also to assess the penetration of quasi-solid electrolyte in the porous matrix of DSSCs. Analysis of charge-transfer resistance and dc resistance of impedance of diffusion of tri-iodide reveals enhanced mobility of tri-iodide in DSSCs. Photovoltaic parameters of quasi-solid DSSCs show an increased open-circuit photovoltage due to the enlarged photoelectrode film porosity and the shift of redox level. Better penetration of quasi-solid electrolyte has a predominant advantage over the negative effect caused by lose of photocurrent, to some extent, as a result of decreased adsorbed dye. The best result of this beneficial outcome occurs when the PEG loading is 20%, giving an overall cell efficiency of 5.1%. (C) 2011 Elsevier B.V. All rights reserved.			
***Metric 435:* stabilized high-effective area conversion efficiency**			
40158479	1997	Kiyama, S.*et al*., (1997). Development of high-efficiency a-Si solar cell submodule with a size of 30 cm x 40 cm, *Solar Energy Materials And Solar Cells*	• 8.64%

[ii] BRAT is a web-based tool for text annotation used to add notes to existing text documents. BRAT is designed in particular for *structured* annotation, where the notes are not freeform text but have a fixed form that can be automatically processed and interpreted by a computer.

Figure 8a: Display of Articles with Specific Performance Metric Claims

Using this tool an analyst can select a documentation citation to see the corresponding document details, as is shown in Figure 8b.



Overview

ID	425304887
Title	*Modification of nanocrystalline porous films by poly(ethyleneglycol) for quasi-solid dye-sensitized solar cells*
Authors	X. Zhao, Q. Tai, S. Xu, B. Chen, H. Hu, B. Sebo
Published	2011
Source	SEEDS.WOS

Abstract

> A convenient way is experimented to reduce the amount of dye in quasi-solid DSSCs but raise open-circuit photovoltage and photocurrent density. AFM stereoscopic morphology and calculated roughness of root mean square indicates looser porous configuration is formed in the modified TiO(2) film which is beneficial for the penetration of quasi-solid electrolyte. Decreased content of sensitized dye is confirmed by UV-vis absorption spectra. Electrochemical impedance spectroscopy is employed to characterize the transport and recombination of electrons and also to assess the penetration of quasi-solid electrolyte in the porous matrix of DSSCs. Analysis of charge-transfer resistance and dc resistance of impedance of diffusion of tri-iodide reveals enhanced mobility of tri-iodide in DSSCs. Photovoltaic parameters of quasi-solid DSSCs show an increased open-circuit photovoltage due to the enlarged photoelectrode film porosity and the shift of redox level. Better penetration of quasi-solid electrolyte has a predominant advantage over the negative effect caused by lose of photocurrent, to some extent, as a result of decreased adsorbed dye. The best result of this beneficial outcome occurs when the PEG loading is 20%, giving an overall cell efficiency of 5.1%. (C) 2011 Elsevier B.V. All rights reserved.

Citations

Title	Year
Guillemoles, J., Minh, C., Koelsch, M., Jolivet, J., & Cassaignon, S. (2004). Electrochemical comparative study of titania (anatase, brookite and rutile) nanoparticles synthesized in aqueous medium, *Thin Solid Films*	2004
Mitate, T., Islam, A., Chiba, Y., Koide, N., & Han, L. (2006). Modeling of an equivalent circuit for dye-sensitized solar cells: improvement of efficiency of dye-sensitized	2006

Figure 8b: Document Entry Accessed from Performance Metric Table

With claims data extracted by Helios, we then conducted a series of post-Helios analyses to see if the extracted data could be used to reproduce NREL's charting of conversion efficiency achievement for each technology, using the following process.

1. All percentage metrics were extracted from DSSC document set
2. Percentage metrics not related to conversion efficiency filtered out
3. Further filtering identified maximum value reported for each year
4. Final filter for annual maximums that are new global maximums

Figure 9a below shows a scatterplot of Helios-extracted percentage metrics for CdTe (step 1). Conversion efficiency is obviously not the only performance metric of interest to researchers. We believe that the data plots well below the top trend line are most likely articles on experiments that attempted to maximize performance along other criteria, such as cell size, manufacturability, or safety. Consultation with Professor Michael Heben at the University of Toledo, an expert in CdTe, supports this assertion. According to Professor Heben, industrial researchers conducted R&D on CdTe during the period from 1995 to 2010 primarily to develop thin-film cells suitable for mass manufacturing.

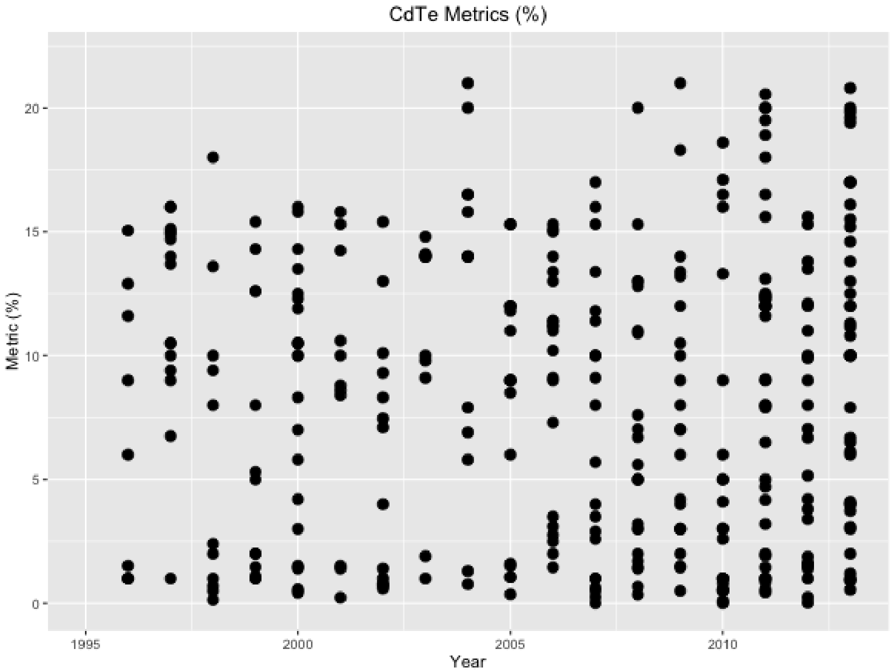


Figure 9a: Extracted CdTe Percentage Metrics (Step 1)

Figure 9b, 9c, and 9d below illustrate how the remaining steps in the process described above converge on a curve that approximates the NREL performance curve. Finally, Figure 9e shows NREL's charting of the maximum achieved CdTe conversion efficiency.

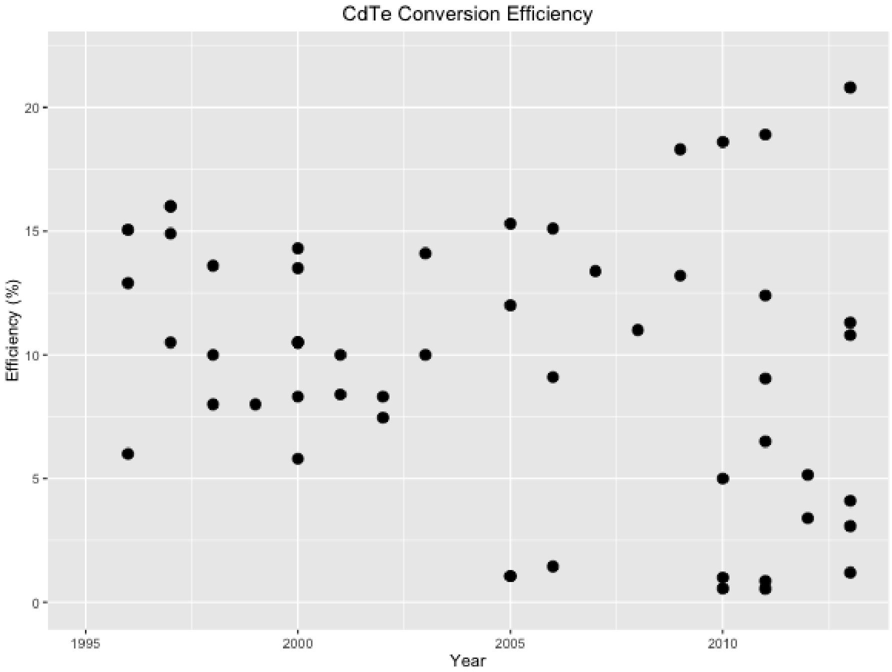


Figure 9b: Extracted CdTe Conversion Efficiency Metrics (Step 2)

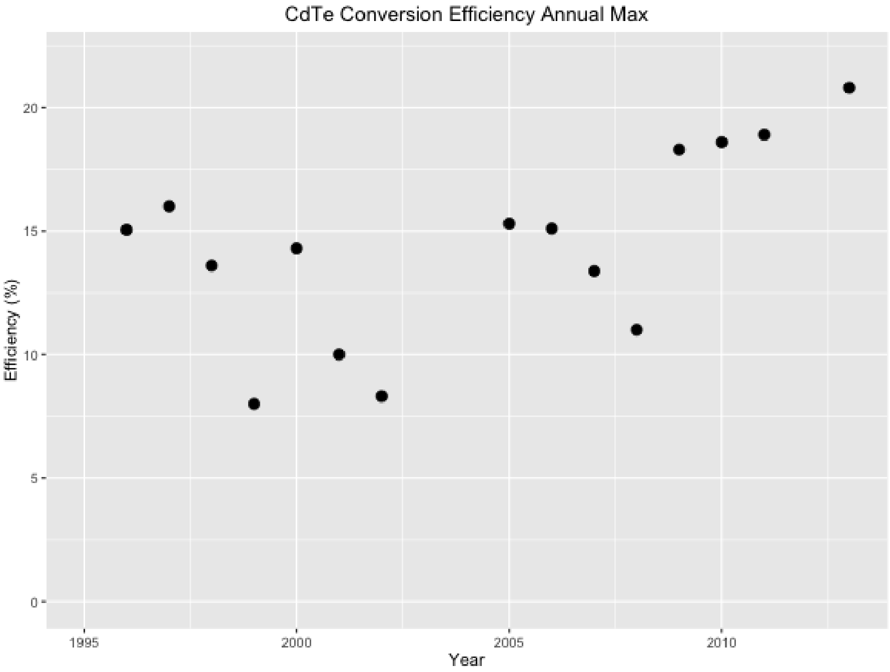


Figure 9c: Annual Maximum Extracted CdTe Efficiency Metrics (Step 3)

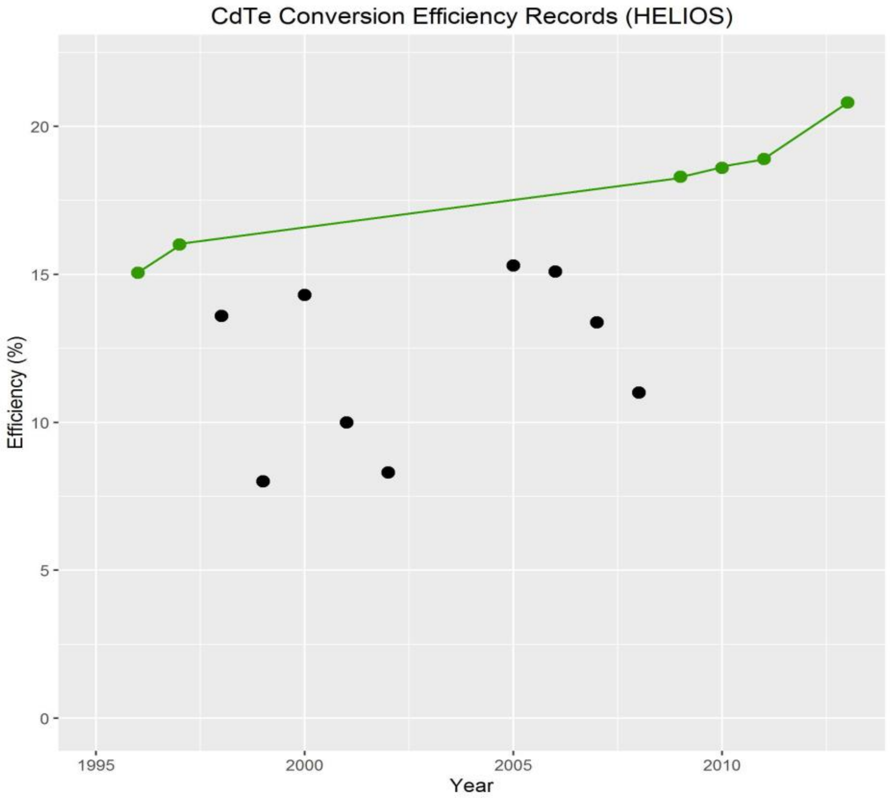


Figure 9d: New Global Maximum Extracted CdTe Efficiency Metrics (Step 4)

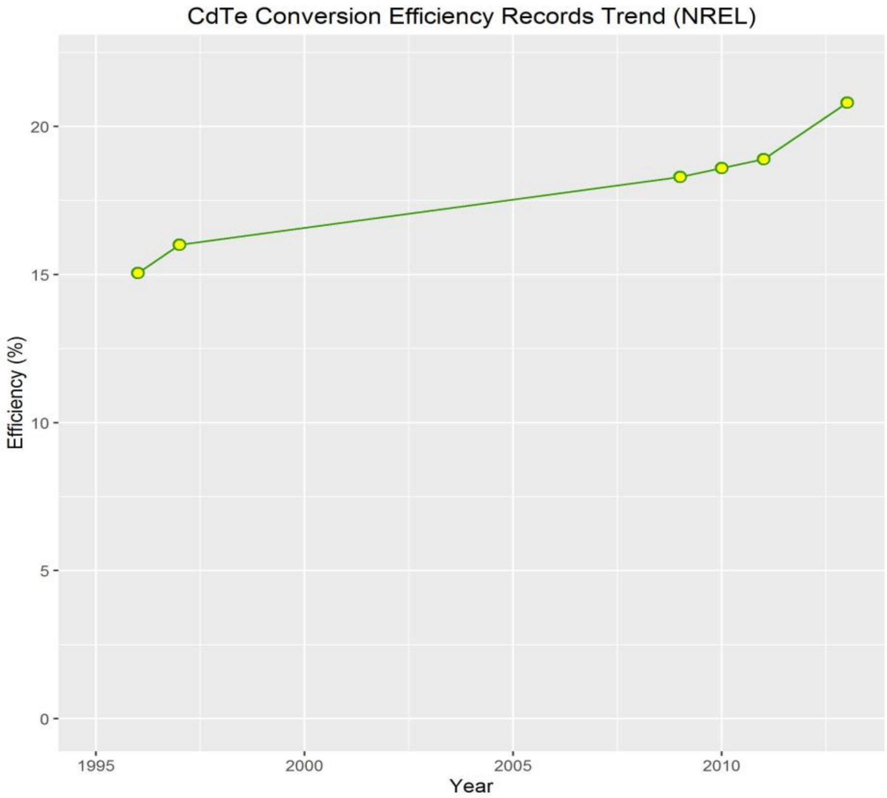


Figure 9e: NREL Conversion Efficiency Record

In the case of CdTe above, the method did a very effective job of matching the NREL chart. In Figures 10a and 10b we repeated the same analysis for DSSC and multijunction technology. In the latter case, the NREL chart has multiple lines corresponding to different number of junctions (2, 3, or 4) and differences in concentrator versus non-concentrator solar cell conversion.[iii] Our analyses did not distinguish number of junctions, though that could be done via the Helios platform using a machine learning approach similar to the one used to extract the performance metrics.

[iii] Concentrator PV systems concentrate terrestrial sunlight many fold onto relatively small cell or panel surface areas in order to generate more energy.

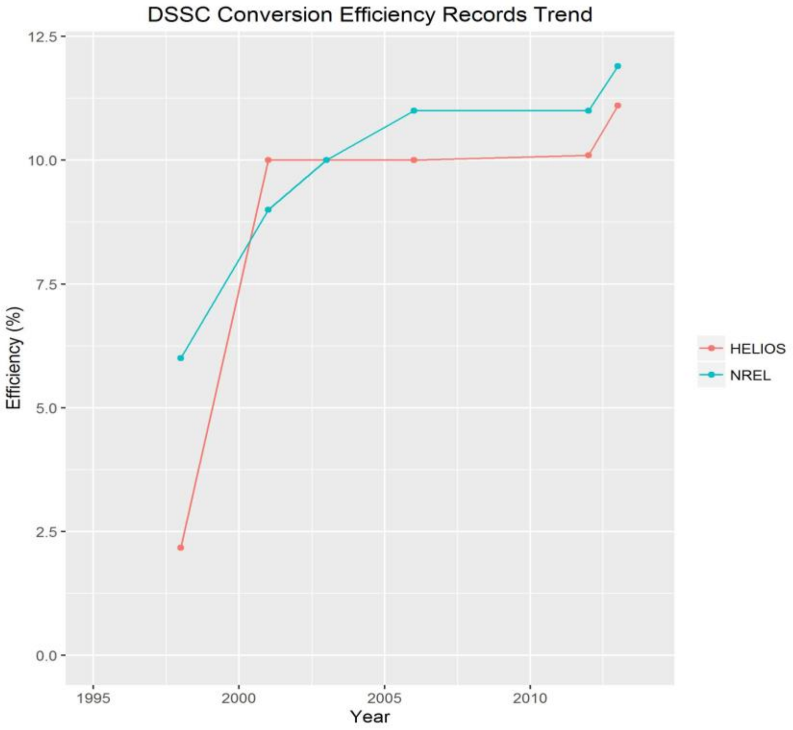


Figure 10a: Helios versus NREL DSSC Conversion Efficiency

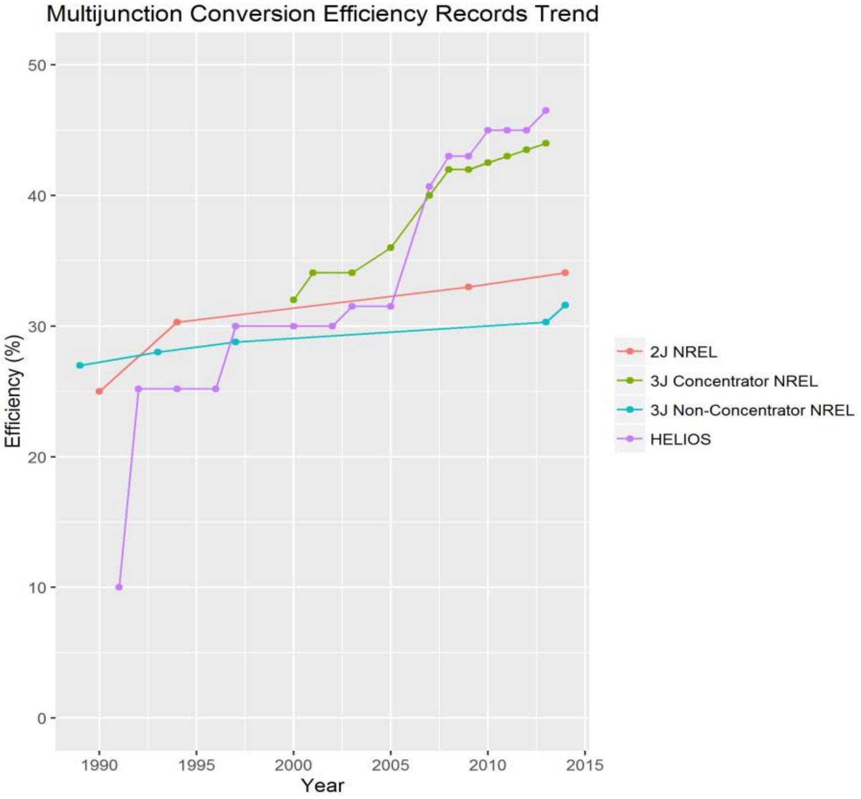


Figure 10b: Helios versus NREL Multi-Junction Conversion Efficiency

The Geography of Photovoltaic Research

Using the articles from the Web of Science document groups, we performed an initial exploration of the relationship between location and patterns of research activity. For each publication, we followed a geocoding process similar to Waltman, Tijssen, & Eck (2011); we extract author affiliation information from the XML records, ignore all but city, state/province and country, and use a geocoding service to convert the address information to geographic coordinates. Addresses were processed using the open source twofishes geocoder to provide city-level precision latitude and longitude coordinates for each author affiliation in the data set (Blackman, 2012).

One issue in geospatial analysis is the somewhat arbitrary placement of political and other administrative boundaries, such as national borders or metropolitan areas. To control for that, we partition the globe into a discrete grid and assign each publication to one or more grid locations based on the spatial intersection of the grid boundaries and the geographic location of the publication's author affiliations. For this analysis we employ the Icosahedral Snyder Equal Area Aperture 3 Hexagon (ISEA3H) Geodesic Discrete Global Grid described by Sahr, White, & Kimerling (2003). A level 8 ISEA3H grid composed of 65,612 cells was generated using the public domain software package DGGRID (Sahr, 2013). Each hexagonal cell corresponds to approximately 7,774 km2 – approximately the same area as a 100 km diameter circle.

We geo-located authors based on their given addresses, or the address of their affiliated institutions. We then explored two aspects of geospatial analysis:

1. Patterns in co-authorship—clustering researchers by their propensity to co-author articles with other particular researchers
2. Patterns in citation—clustering researchers by their propensity to cite the same references in their articles

Figures 11 below show the results. Each community is plotted with a unique color.

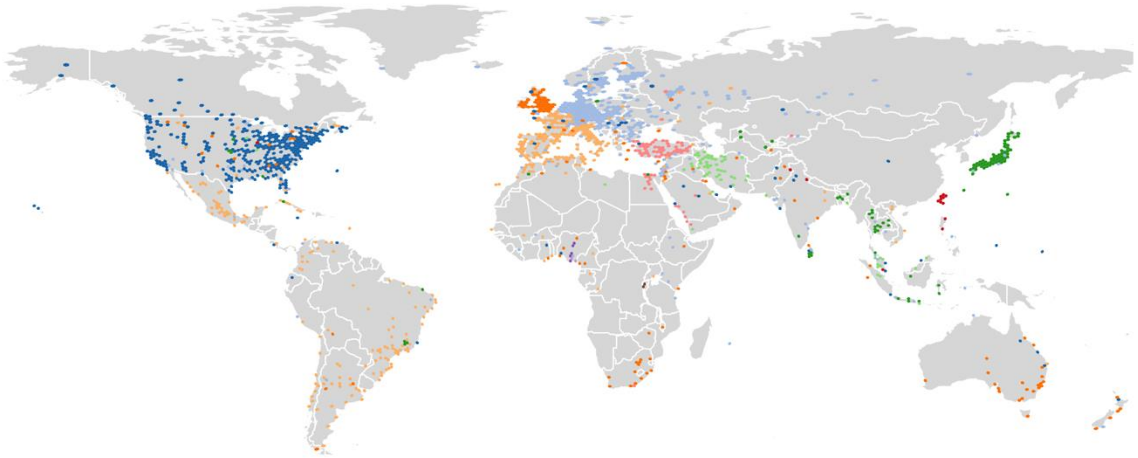


Figure 11a: Co-authorship Community Detection (colors = distinct communities)

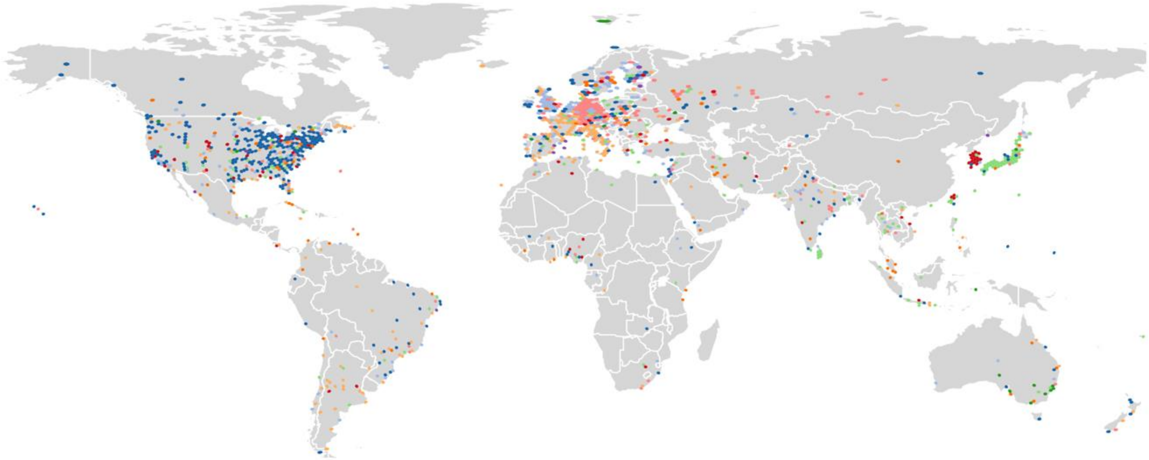


Figure 11b: Citation Community Detection (colors = distinct communities)

Despite the globalization of scientific knowledge, co-authorship networks still tend to be geographically concentrated. Even in the information age, it is easier to collaboration on an article with a colleague who is easily accessible in person, rather than one who is remote. The citation community chart also shows local preferences, with researchers in particular locations having a tendency to cite the same bodies of work, but the localization appears less acute than with co-authorship (this assessment is via visual inspection – more formal measures of geographic concentration could be explored in future work). Less locally concentrated citation makes sense in that the broader community shares a common recognition of the most important and most recent discoveries upon to which their work relates and upon which it builds. Even with citation, however, we observe some location, especially in places where language it more likely to be a barrier, such as in Japan.

Subtask 4.6: Analysis Team Support

The Analysis team provided ongoing support during the project.

Task 5: Platform Validation and Deployment

Subtask 5.1: Perform In-Depth Helios-Aided Exploration of Case Study Topics

The Technology Analysis Team identified specific individual “breakthroughs” in each of the three technologies for later analysis. We developed a set of bibliometric and topic modeling related metrics that could be used to help analysts assemble breakthrough narratives that identified the researchers responsible for a given breakthrough, the researchers whose research they were following, the topics they were working on.

To test the approach, we constrained the breakthroughs used as targets with the following criteria:

- Timeframe. XML-formatted records from USPTO ingested into Helios only date back to approximately 2004. Also, in the scientific literature, Web of Science records only include abstracts starting from the early 1990s. Therefore, text analysis would be much sparser for earlier breakthroughs.
- Document availability. Ideally, the breakthrough would be described by a single source document, or a small set of source documents, found in the Helios database. This would enable more explicit tracking of individuals, institutions, and topics from that breakthrough event to earlier developments in science and technology.
- Availability of scientific “tracing” evidence. The Helios approach assumes that a given breakthrough is based on an accumulation of prior discoveries and inventions in a variety of fields, which are then combined with novel sets of concepts to find a method for overcoming a particular barrier to increased performance. As a result, we believe the approach will provide the best findings when we look at breakthroughs in recent years, when much of the literature that references or is related to that breakthrough is also available in the Helios database.

The team identified significant breakthroughs in each technology’s history as indicated by a substantial jump in efficiency and used the document describing each breakthrough as the “seed” for constructing case document groups from each dataset (primarily WoS and USPTO). For example, Japanese researchers achieved a breakthrough in DSSC technology in 2006, managing to build a DSSC with an efficiency of 11% (Chiba et al, 2006); see Table 9.

Technology group	DSSC
Date of breakthrough	07/2006 (published)
Docs associated with breakthrough	Chiba, Y, Islam, A, Watanabe, Y, Komiya, R, Koide, N, Han, L (2006). Dye-sensitized solar cells with conversion efficiency of 11.1%. *Japanese Journal of Applied Physics*, **45**(25), L638-L640.
Institutions associated with breakthrough	New Technology Development Center, Solar Systems Development Group, Sharp Corporation (Chiba, Watanabe, Komiya, & Koide) National Institute of Material Sciences, Tsukuba, Japan (Islam & Han)
Approximate timeframe of interest	1990 to 2009

Table 9: DSSC Breakthrough Document Group

In the breakthrough study, the team established the evolutionary history of the breakthrough technology for eventual comparison with the Helios exploration, focusing on manually identifying key citations, authors, institutions, and ideas related to the technological breakthrough.

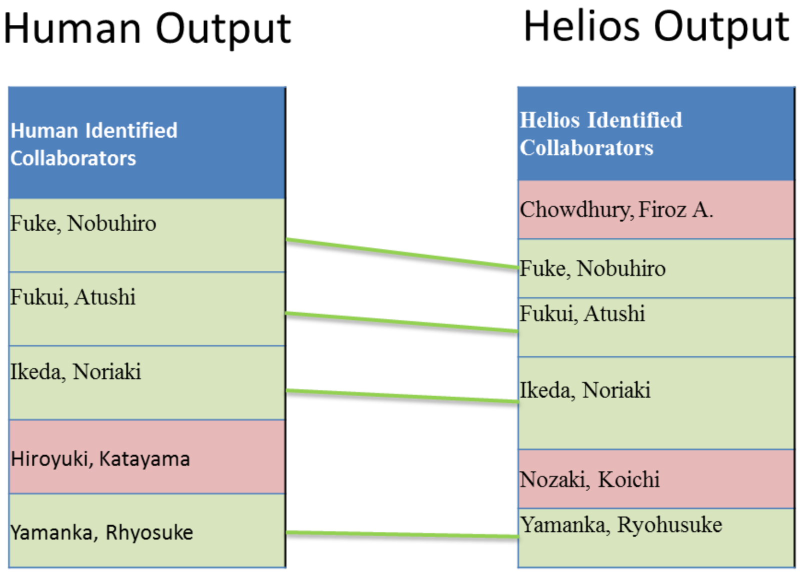
Subtask 5.2: Report and Validate Results

Following the development of the breakthrough case studies in Subtask 5.1, we used Helios to perform seven analyses on each technological document group, as described in Table 10 below. These analyses are meant to identify with whom and on which topics the breakthrough innovators were working during the period of the breakthrough and the topics they were working on. The measure names are arbitrary. Ideally, the output associated with these measures would be consistent with the human-generated breakthrough case studies.

Measure	Entity analyzed	Analytical technique
W1	Most prominent collaborators with co-authors	PageRank on co-authorship network for all members of the team
W2	Key works cited by members of the research team	PageRank on citations in the (backward looking) citation network of the research team
W3	Key topics from breakthrough document group	Topic model on all works authored by team members and their citation network prior to breakthrough document
W4	Key topics from expanded breakthrough document group	Topic model on co-author network of team members
W5	Key topics of works cited by research team members prior to breakthrough	Topic model on the citation (backward looking) network for the team members
W6	Key topics from technology group	Topic model on each technology dataset over period of interest

Table 10: Helios Analytical Approach

The results indicate that the Helios platform is capable of identifying the researchers responsible for a given breakthrough, the immediate network of those researchers, and several of the topics they were working on around the time of the breakthrough. The human analyst identified five key collaborators of the breakthrough team, while Helios identified six, four of whom overlap with those identified by the human analyst; see Figure 12.



Human Output

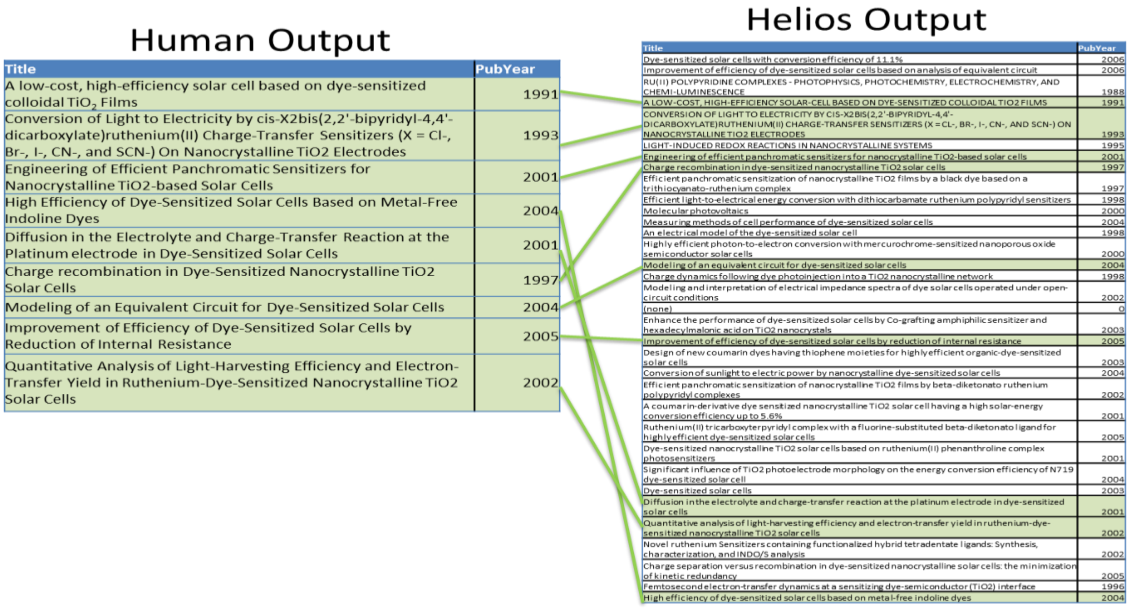
Human Identified Collaborators
Fuke, Nobuhiro
Fukui, Atushi
Ikeda, Noriaki
Hiroyuki, Katayama
Yamanka, Rhyosuke

Helios Output

Helios Identified Collaborators
Chowdhury, Firoz A.
Fuke, Nobuhiro
Fukui, Atushi
Ikeda, Noriaki
Nozaki, Koichi
Yamanka, Ryohusuke

Figure 12: Helios Identification of Key Collaborators (versus Human Analyst)

Helios was also able to accurately identify all of the important and relevant papers in the literature that formed the basis of the science that preceded the breakthrough. The human analyst identified nine such key papers. Helios identified all of these, though was less discriminant than the human analyst, identifying a pool of 39 papers, as shown in Figure 13.



Human Output

Title	PubYear
A low-cost, high-efficiency solar cell based on dye-sensitized colloidal TiO$_2$ Films	1991
Conversion of Light to Electricity by cis-X2bis(2,2'-bipyridyl-4,4'-dicarboxylate)ruthenium(II) Charge-Transfer Sensitizers (X = Cl-, Br-, I-, CN-, and SCN-) On Nanocrystalline TiO2 Electrodes	1993
Engineering of Efficient Panchromatic Sensitizers for Nanocrystalline TiO2-based Solar Cells	2001
High Efficiency of Dye-Sensitized Solar Cells Based on Metal-Free Indoline Dyes	2004
Diffusion in the Electrolyte and Charge-Transfer Reaction at the Platinum electrode in Dye-Sensitized Solar Cells	2001
Charge recombination in Dye-Sensitized Nanocrystalline TiO2 Solar Cells	1997
Modeling of an Equivalent Circuit for Dye-Sensitized Solar Cells	2004
Improvement of Efficiency of Dye-Sensitized Solar Cells by Reduction of Internal Resistance	2005
Quantitative Analysis of Light-Harvesting Efficiency and Electron-Transfer Yield in Ruthenium-Dye-Sensitized Nanocrystalline TiO2 Solar Cells	2002

Helios Output

Title	PubYear
Dye-sensitized solar cells with conversion efficiency of 11.1%	2006
Improvement of efficiency of dye-sensitized solar cells based on analysis of equivalent circuit	2006
RU(II) POLYPYRIDINE COMPLEXES - PHOTOPHYSICS, PHOTOCHEMISTRY, ELECTROCHEMISTRY, AND CHEMI-LUMINESCENCE	1988
A LOW-COST, HIGH-EFFICIENCY SOLAR-CELL BASED ON DYE-SENSITIZED COLLOIDAL TIO2 FILMS	1991
CONVERSION OF LIGHT TO ELECTRICITY BY CIS-X2BIS(2,2'-BIPYRIDYL-4,4'-DICARBOXYLATE)RUTHENIUM(II) CHARGE-TRANSFER SENSITIZERS (X = CL-, BR-, I-, CN-, AND SCN-) ON NANOCRYSTALLINE TIO2 ELECTRODES	1993
LIGHT-INDUCED REDOX REACTIONS IN NANOCRYSTALLINE SYSTEMS	1995
Engineering of efficient panchromatic sensitizers for nanocrystalline TiO2-based solar cells	2001
Charge recombination in dye-sensitized nanocrystalline TiO2 solar cells	1997
Efficient panchromatic sensitization of nanocrystalline TiO2 films by a black dye based on a trithiocyanato-ruthenium complex	1997
Efficient light-to-electrical energy conversion with dithiocarbamate ruthenium polypyridyl sensitizers	1998
Molecular photovoltaics	2000
Measuring methods of cell performance of dye-sensitized solar cells	2004
An electrical model of the dye-sensitized solar cell	1998
Highly efficient photon-to-electron conversion with mercurochrome-sensitized nanoporous oxide semiconductor solar cells	2000
Modeling of an equivalent circuit for dye-sensitized solar cells	2004
Charge dynamics following dye photoinjection into a TiO2 nanocrystalline network	1998
Modeling and interpretation of electrical impedance spectra of dye solar cells operated under open-circuit conditions	2002
(none)	0
Enhance the performance of dye-sensitized solar cells by Co-grafting amphiphilic sensitizer and hexadecylmalonic acid on TiO2 nanocrystals	2003
Improvement of efficiency of dye-sensitized solar cells by reduction of internal resistance	2005
Design of new coumarin dyes having thiophene moieties for highly efficient organic-dye-sensitized solar cells	2003
Conversion of sunlight to electric power by nanocrystalline dye-sensitized solar cells	2004
Efficient panchromatic sensitization of nanocrystalline TiO2 films by beta-diketonato ruthenium polypyridyl complexes	2002
A coumarin-derivative dye sensitized nanocrystalline TiO2 solar cell having a high solar-energy conversion efficiency up to 5.6%	2001
Ruthenium(II) tricarboxyterpyridyl complex with a fluorine-substituted beta-diketonato ligand for highly efficient dye-sensitized solar cells	2005
Dye-sensitized nanocrystalline TiO2 solar cells based on ruthenium(II) phenanthroline complex photosensitizers	2001
Significant influence of TiO2 photoelectrode morphology on the energy conversion efficiency of N719 dye-sensitized solar cell	2004
Dye-sensitized solar cells	2003
Diffusion in the electrolyte and charge-transfer reaction at the platinum electrode in dye-sensitized solar cells	2001
Quantitative analysis of light-harvesting efficiency and electron-transfer yield in ruthenium-dye-sensitized nanocrystalline TiO2 solar cells	2002
Novel ruthenium Sensitizers containing functionalized hybrid tetradentate ligands: Synthesis, characterization, and INDO/S analysis	2002
Charge separation versus recombination in dye-sensitized nanocrystalline solar cells: the minimization of kinetic redundancy	2005
Femtosecond electron-transfer dynamics at a sensitizing dye-semiconductor (TiO2) interface	1996
High efficiency of dye-sensitized solar cells based on metal-free indoline dyes	2004

Figure 13: Helios Identification of Key Papers (versus Human Analyst)

Through topic modelling, Helios was able to identify the major technological accomplishments that led up to the breakthrough and some of the key concepts associated with the breakthrough itself. Topic models for the whole technology group were created (W6 in Table 10), and then the model was restricted to just the documents in the expanded breakthrough group (W3 through W5 in Table 10). These topic models were then scored according to how much they differentiate from the topic model of the whole technology group, providing a picture of the breakthrough document's novel contributions to the evolution of the technology.

Topic	Score	Most Frequent Terms
21	4521	2, 4, **ru**, complexes, ii, **ruthenium**, complex, bpy, ncs, **bipyridine**, 1, nanocrystalline, **tio2**, 3, ligand, ligands, dcbpy, l, h, sensitizers, **polypyridyl**, x, cn, cells, solar, pf6, films, **terpyridine**, cl, sensitizer
35	2848	cells, charge, solar, photovoltaic, light, films, semiconductor, transition, absorption, **nanocrystalline**, band, conversion, large, dye, oxide, devices, **transition metal**, current, electric, conventional, **sensitized**, molecular, light absorption, present, separation, carrier, new, metal, cell, carrier transport
27	1422	2, circuit, cm, ma, short, ma cm, solar, open, **open circuit**, **v**, **sc**, **density**, cell, voc, photocurrent, current, **fill**, voltage, **factor**, j, **short circuit photocurrent**, **efficiency**, short **circuit current**, m, mv, cells, conversion, ff, performance, i

Table 11: Helios Generated Topic Models of Concepts in Breakthrough Group

The top three topics from the resulting output are shown above in Table 11. The bolded terms were identified as critical concepts in the breakthrough document group in the human-generated case study. These results indicate that when combined with claims analysis, the Helios platform could potentially be used to effectively identify significant breakthroughs and the important researchers and ideas associated with those breakthroughs.

The methodology described here potentially could be improved by dividing the analysis into three time periods: documents leading up to the breakthrough, documents associated with the breakthrough and occurring simultaneously, and documents that follow after the breakthrough. This division would likely present a more accurate picture of the history of research, the novel elements of each breakthrough, and the forward-facing impact that breakthroughs had on research and technology.

Task 6: Method Expansion

Subtask 6.1: Expansion Beyond Small Set

During analyst use of the Helios platform, machine-based topic modeling would highlight terms a domain-knowledgeable analyst would be able to identify with a specific technology concept, highlighting that concept as something meriting further investigation within Helios, provided sufficient data were available to explore this new topic within Helios.

In pushing our analysis past the initial set of technologies of CdTe, DSSC, and multijunction, we wanted to rely on this capability of Helios for two reasons. First, it would demonstrate the capability of Helios to highlight potentially new topics. Second, a topic highlighted within the existing technology document sets would be more likely to have some coverage in our existing data, as the project budget did not allow for additional data collection.

When reviewing the topic modeling of the three core technology groups, it appeared that the terms of DSSC Topic 40 (see Table 12 below) were broadly consistent with perovskite. This then represented a promising test subject given our data limitations. Perovskite PV research came out of DSSC research, and so there was a reasonable likelihood that there was a critical mass of perovskite papers in our data set.

oxide metal oxides metal oxide iron nickel nio metal oxides zro2 mixed mgo	iron oxide nitride titanium oxide titanium oxide films vanadium solid binary cobalt fe2o3 hematite	alpha-fe2o3 oxide nanoparticles niobium manganese cerium metals ternary aluminum perovskite nickel oxide ceo2

Table 12: Terms of DSSC Topic 40

We created a mini perovskite case study, identifying key researchers, institutions, and documents associated with the technology's evolution, and looked to see if the applicable papers were in the DSSC dataset. Unfortunately, the data obtained for this project was obtained in 2014, the most recent papers in the dataset are from 2013. Given the recency of most solar cell research on perovskite, only one of the documents from the mini case study was included in the existing SEEDS dataset. We were therefore unable to run expanded analysis for perovskite without another large data purchase from Thomson Reuters.

Conclusions

Project Contributions and Final Deliverables

During this project we were able to demonstrate the use of text analytics, in particular topic modeling, to identify important shifts in the development of photovoltaic technology. In addition to identifying the technical changes themselves, we were able to highlight important contextual elements associated with these developments, in particular the researchers responsible for each development, and to a more limited extent, the set of collaborators these innovators worked with and the topics they were exploring ahead of each development.

As a proof-of-concept project, a number of approaches and capabilities were demonstrated, many at a level of accuracy relative to what a human analyst could do that indicates these initial efforts merit further research. The approaches developed were embedded in a tool that allowed for the interrogation of thousands – and in principle millions – of documents of different kinds. No single analyst or team of analysts could hope to evaluate this much data.

Performance relative to ground truth was assessed primarily through completion of three detailed human-generated case studies of solar cell technology development, along with more targeted human analysis. The performance of the tool and its assessment relative to ground truth, possible refinements to the work, and its prospective use have been written up in detail in this report, which constitutes **Final Deliverable 1**.

With refinements, several of our preliminary approaches seem likely of achieving a level of accuracy and insight that would make them quite valuable in analyzing the evolution of technology across very large corpora of documents. Such a tool would be invaluable to policy makers and R&D managers as both groups seek to understand patterns in the evolution of technology that might shed light on how R&D efforts could be better organized and investment more effectively made.

A Virtual Machine (VM) has been configured to run on a virtualbox host system for delivery to the Department of Energy. The host system can be any OS, and has Ubuntu installed with all necessary supporting tools (postgresql, spark, jooq, etc.). Scripts used to derive results during the SEEDS project are included, and a command line interface is available to interact with the Helios analytic engine. Documentation is loaded into a doc directory on the VM with the software. Provision of this prototype machine (virtual machines), which gives the Government unlimited use rights, constitutes Final Deliverable 3.

During the course of the research, project team members disseminated emerging results in appropriate forums, which constitute Final Deliverable 2 as follows:

Journal Articles

Full Author List	"Article Title"	Journal Name	Volume Number	No #	pp. (##-##)
Jeffrey M. Alexander, John Byrnes, John Chase, Christina Freyman	Text Analytics for Retrospective Technology Tracing: The Helios Project on Photovoltaics Innovation	*Technological Forecasting & Social Change (submitted)*			
John Byrnes, John Chase, Christina Freyman, Lucien Randazzese	Automated Technical Achievement Claims Extraction from Scientific Literature	*currently searching for appropriate outlet*			

Conference Publications

Full Author List	"Article Title"	Conference/Proceedings Title	Conference Location	Dates	Paper Number
Chase, John J. & Cunningham, Scott W.	"Putting citation in its place: Exploring the role of geography in publication impact."	*Proceedings of the 2014 STI Conference on Science & Technology Indicators*	Leiden, the Netherlands	3 to 5 Sept, 2014	
Alexander, Jeffrey M., Byrnes, John, Chase, John, & Freyman, Christina	"Text Analytics for Retrospective Technology Tracing"	2015 Atlanta Conference on Science & Innovation Policy	Atlanta, GA	17 to 19 Sept, 2015	249

Conference Presentations

Full Author List	"Paper Title"	Session/Symposium/Conference	Conference Location	Dates	Paper Number
Chase, John, J., Alexander, Jeffrey M., Byrnes, John & Freyman, Christina	"Understanding Solar PV Evolution through Text Analytics" (Poster)	2013 Atlanta Conference on Science & Innovation Policy	Atlanta, GA	26-28 Sept, 2013	
Chase, John, J., Alexander, Jeffrey M., Byrnes, John & Freyman, Christina	"Understanding Solar PV Evolution through Text Analytics" (Poster)	FEDLINK Symposium on Analytical Methods for Technology Forecasting	Washington, DC	6 March 2014	
Chase, John, J., Alexander, Jeffrey M., Byrnes, John & Freyman, Christina	"HELIOS: Understanding Solar PV Evolution through Text Analytics" (Poster)	2014 SunShot Summit	Anaheim, CA	20-21 May 2014	2136
Alexander, Jeffrey M., Byrnes, John, Chase, John, & Freyman, Christina	"Text Analytics for Retrospective Technology Tracing"	2015 Atlanta Conference on Science & Innovation Policy	Atlanta, GA	17 to 19 Sept, 2015	249

Milestone Achievement

Task 1 Milestone

Based on our review of available data sources and extended discussion of the planned work with the project advisory board, we developed and submitted a revised project plan.

Task 2 Milestone

Three detailed case studies were completed to serve as baseline for later project efforts. The case studies highlighted numerous technical and historical aspects of the evolution of the three focus technologies (CdTe, DSSC, multijunction), but the project team did not believe they contained sufficient new-to-world insight to merit writing a publication for peer review.

Task 3 Milestone

The Helios platform generated visualizations both of coclustering and of topic modeling that were reviewed by the analysis team. Figures 1, 2, 4, 7a, 7b and Tables 11 and 12 were produced by the Helios tool for inspection of topic models and coclustering results. The Helios platform generates citation, co-citation, and co-authorship networks and computes network centrality scores for papers and authors in these networks. The citation graph in Figure 3 and centrality scores in Table 5 are examples of this output from Helios. These outputs were reviewed by the Analysis team as part of their study of photovoltaic emergence.

Go/No-Go Decision Point

The team submitted a continuation report on May 16, 2014, after the first 15 months of the project. Following submission of this report and presentation of the work of Period 1 (originally scheduled for 12 months but extended to 15 months), DOE elected to fund the second period of work for the project.

Task 4 Milestone

The Helios platform was developed into a stand-alone system that runs in a Virtual Machine. The system supports topic modeling and network influence analysis as described above. These analyses run across the ingested data sets.

Task 5 Milestone

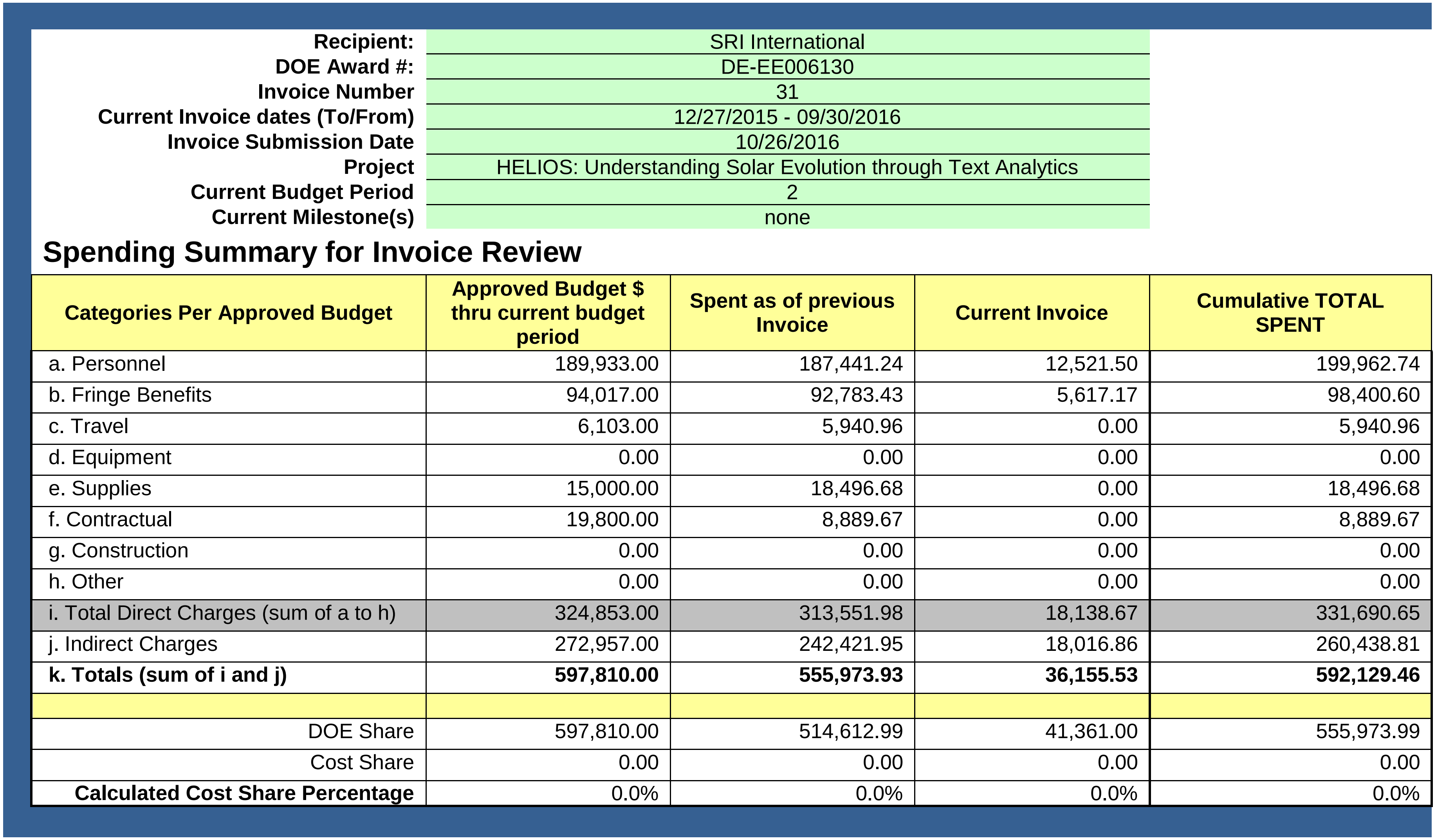
We were able to develop a preliminary methodology by which the specific researchers and topics of scientific inquiry associated with step-function changes in technical performance, what we call breakthroughs, were identified by Helios.

Task 6 Milestone

Given the limitation of the project data, we were unable to test Helios performance on technologies outside the initial set.

Budget and Schedule

Budget information for the project follows:



Recipient:	SRI International
DOE Award #:	DE-EE006130
Invoice Number	31
Current Invoice dates (To/From)	12/27/2015 - 09/30/2016
Invoice Submission Date	10/26/2016
Project	HELIOS: Understanding Solar Evolution through Text Analytics
Current Budget Period	2
Current Milestone(s)	none

Spending Summary for Invoice Review

Categories Per Approved Budget	Approved Budget $ thru current budget period	Spent as of previous Invoice	Current Invoice	Cumulative TOTAL SPENT
a. Personnel	189,933.00	187,441.24	12,521.50	199,962.74
b. Fringe Benefits	94,017.00	92,783.43	5,617.17	98,400.60
c. Travel	6,103.00	5,940.96	0.00	5,940.96
d. Equipment	0.00	0.00	0.00	0.00
e. Supplies	15,000.00	18,496.68	0.00	18,496.68
f. Contractual	19,800.00	8,889.67	0.00	8,889.67
g. Construction	0.00	0.00	0.00	0.00
h. Other	0.00	0.00	0.00	0.00
i. Total Direct Charges (sum of a to h)	324,853.00	313,551.98	18,138.67	331,690.65
j. Indirect Charges	272,957.00	242,421.95	18,016.86	260,438.81
k. Totals (sum of i and j)	**597,810.00**	**555,973.93**	**36,155.53**	**592,129.46**
DOE Share	597,810.00	514,612.99	41,361.00	555,973.99
Cost Share	0.00	0.00	0.00	0.00
Calculated Cost Share Percentage	0.0%	0.0%	0.0%	0.0%

The project work plan originally extended from April, 1, 2013 to March 31, 2014 for Period 1 and April 1, 2014 to March 31, 2015 for Period 2. There were seven modifications to the project. These extended Period 1 to June 30, 2014 and Period 2 to September 30, 2016, and also accounted for the change in PI from Jeffrey Alexander to Lucien Randazzese. There was a single go/no-go decision after period one which ended on June 30, 2014.

Path Forward

Under the SEEDS II program, the team intends to extend the Helios technology to identify Technology Readiness Level for solar technologies based on publications and reports. We will extend entity recognition tools that were developed under the Helios project in order to recognize techniques and applications.

We are also extending these technology recognition tools under an NREL contract to recognize Critical Materials for the Department of Energy. We will recognize both material mentions and their applications in an effort to identify new emerging applications for critical materials and to recognize alternative materials being developed for existing applications.

We continue to develop Helios for IARPA and to develop applications of it for Horizon Scanning (identifying new emerging technical capabilities) for the intelligence community and for retrospective analysis of technology development for research funders.

References

Bettancourt, L. and Kaiser, D. I. (2015). “Formation of Scientific Fields as a Universal Topological Transition.” SFI Working Paper 2015-03-009.

Blackman, D. (2012). Twofishes. GitHub Repository. Retrieved from https://github.com/foursquare/twofishes.

Chandler, D.L. (2015). New kind of “tandem” solar cell developed. Press release, MIT: Cambridge, MA, 24 March. Retrieved from http://newsoffice.mit.edu/2015/tandem-solar-cell-0324.

Chiba, Y, Islam, A, Watanabe, Y, Komiya, R, Koide, N, Han, L (2006). “Dye-sensitized solar cells with conversion efficiency of 11.1%.” *Japanese Journal of Applied Physics*, 45(25), L638-L640.

Gui, J., Yao, C.-Q., Zeng, W., and Zhang, Z.-F. (2015). How does basic research promote the innovation for patented invention: a measuring of NPC and technology coupling. Proceedings of the 2015 International Conference on Management Science and Management Innovation.

Leydesdorff, L., Alkemade, F., Heimerks, G., & Hoekstra, R. (2014). Geographic and Technological Perspectives on “Photovoltaic Cells:” Patents as Instruments for Exploring Innovation Dynamics. CoRR.

Luan, C., Liu, Z., and Wang, X. (2013). “Divergence and convergence: technology-relatedness evolution in the solar energy industry.” Scientometrics, 97, 461—475.

McKeown, K., Daume, H., Chaturvedi, S., Paparrizos, J., Thadani, K., Barrio, P., Biran, O., Bothe, S., Collins, M., Fleischmann, K., Gravano, L., Jha, R., King, B., McInerney, K., Moon, T., Neelakantatan, A., O’Seaghdha, D., Radev, D., Templeton, C., and Teufel, S. (2016). Predicting the impact of scientific concepts using full-text features. Journal of the Association for Information Science and Technology, forthcoming.

Packalen, M. and Bhattacharya, J. (2012). “Words in Patents: Research Inputs and the Value of Innovativeness in Invention.” NBER Working Paper No. 18494, National Bureau for Economic Research, Cambridge, MA.

Sahr, K. (2013). DGGRID (version 6.1). Retrieved from http://discreteglobalgrids.org/software.

Sahr, Kevin, Denis White, and A. Jon Kimerling. "Geodesic discrete global grid systems." Cartography and Geographic Information Science 30.2 (2003): 121-134.

Small, Boyak, Klavans. (2014). Identifying emerging topics in science and technology. Research Policy.

Strumsky, D. and Lobo, J. Identifying the sources of technological novelty in the process of innovation. Research Policy, 44, 1445-1461.

Rai, V., Metteauer, M., Querejazu, D., Wise, R., & Hamilton, G. (2013). Demand-Pull and Innovation in the US Solar Market. Available at SSRN 2297140.

Tsai, C., Kundu, G., and Roth, D. (2013). “Concept-based analysis of scientific literature.” ACM International Conference on Information and Knowledge Management (CIKM ’13), October 27-November 1, San Francisco, CA.

Venugopalan, S., & Rai, V. (2015). Topic based classification and pattern identification in patents. Technological Forecasting and Social Change, 94, pp. 236-250.

Waltman, L., Tijssen, R. J. W., & Eck, N. J. Van. (2011). Globalisation of science in kilometres. Retrieved February 04, 2014, from http://arxiv.org/pdf/1103.3648v2.pdf.

Zhou, X., Zhang, Y., Porter, A. L., Guo, Y., & Zhu, D. (2014). A patent analysis method to trace technology evolutionary pathways. Scientometrics, 100(3), pp. 705-721.