The Science Library: Curation and visualization of a science gateway repository

Dave Braines¹,², Jane Stockdill-Mander¹, Eunjin Lee³

¹IBM Research Europe, Winchester, UK
²Crime and Security Research Institute, Cardiff University, Cardiff, UK
³IBM, Winchester, UK

Correspondence
Dave Braines, IBM Research Europe, Hursley Park, Winchester, Hampshire SO21 2JN, UK.
Email: dave_braines@uk.ibm.com

Funding information
U.S. Army Research Laboratory and the UK Ministry of Defence, Grant/Award Number: W911NF-16-3-0001

The Science Library is a publicly available, user-centric, interactive portal that we have created to visualize and curate a science gateway repository. It is designed to convey the underlying threads and relationships among scientific publications from a group or consortium, beyond the traditional information captured through references and citations.

Scientific publications from a group or consortium often form a coherent larger body of work with underlying threads and relationships. Rich social, structural, and topical networks between authors and organizations can be identified, and to convey these we have created the publicly available “Science Library” as a user-centric, interactive portal. A key consideration in this endeavor is rapid and efficient curation of the corpus of publications, both in terms of assuring quality, as well minimizing the effort required. For this to be sustainable it must offer substantial benefits to the community and avoid excessive operational cost through cumbersome or complex processes. We describe the agility of the Science Library implementation as a controlled natural language (CNL) semantic knowledge graph and describe the different roles within the community to ensure efficient curation, validation, and provenance of the content. By describing the process of curation and validation, alongside the CNL-based definition of the model we show how relatively non-technical users are able to interact with, and contribute to the Science Library. This provides an extensible approach, initially based around digital library and virtual community capabilities, that can be applied more broadly to support other desired capabilities of Science Gateways.

KEYWORDS
Controlled English, human-computer interaction, science gateways, Science Library, visualization

INTRODUCTION

When publishing a broad body of scientific research, we believe that it is important to retain some of the contextual information about the relationships between the papers, authors, and topics. This helps the reader to understand the "bigger picture" behind each paper, beyond the traditional information captured through references and citations. The International Technology Alliance in Distributed Analytics and Information Science (DAIS-ITA)¹ is a research program conducting fundamental research across a consortium of organizations from academia, industry and government in the US and UK. For this program, and the previous Network and Information Science (NIS) ITA,² we created a large long-running multi-disciplinary research team from multiple organizations who are encouraged to pursue collaborative research across organizational, national, and disciplinary boundaries. In this context, recording the organizational, topical, and social context of the papers is a key consideration for us in quantifying broader impacts and effects of the work.

This work is an extended version of an earlier conference paper³ with most of the additional material provided in Sections 3 and 4; however, the original material has been condensed where possible and all screenshots have been regenerated from the latest running DAIS ITA Science Library.
using the previous conference paper as the example for screenshots. This paper is intended as a stand-alone extension and replacement for the previous work, however due to the necessary revisions, consolidation, and replacement of figures some readers may wish to refer back to the original publication\(^1\) for those details. In addition to covering the various interactive visualizations that enable Science Library visitors to explore the rich knowledge graph of research publication information, we delve more deeply into the processes by which we curate the underlying information for publishing. We believe that the approach we have taken supports the idea of cognitively mediated research discovery\(^2\) both in terms of the interactive visualization and search capabilities provided, but also in the way in which the underlying information is modeled, captured, and shared. We expose multiple perspectives by which a visitor might explore a particular research domain; able to follow different pathways between papers and authors to navigate the cognitive space.

The purpose of this paper is to describe the potential for a highly extensible knowledge curation platform that can be easily extended to support additional concepts, relationships, and information sources without the need for deep technical skills. This paper aims to provide a detailed worked example of an initial attempt to achieve this goal, based on practical experience, and describes in detail the technical implementation of such a solution and the evolving understanding of how to curate and validate the underlying data. This paper aims to provide insight through example, and showcase the underlying technologies and their patterns of usage, with the intent of informing the reader of the specific details of our solution, and the potential for other similar solutions using the same approaches. More specifically: the intent of the paper is to provide rich technical details of our example implementation and to showcase the potential value of a controlled natural language (CNL) based extensible approach, with the detailed worked example to provide concrete support for a variety of user perspectives.

To achieve this in a realistic time frame we use the advantages available when developing such semantic knowledge graphs using a CNL\(^5\), the implementation of which is described in this paper. We also outline the high-level architecture of the Science Library system, and the techniques we use to solicit publication information from the DAIS ITA community. The scientific goals of this work are two-fold: first to capture and convey scientific information in a useful collaborative context, and second, to exercise the English-like language for easily defining and extending complex semantic knowledge graphs without deep technical expertise. The embodiment of the science library application is the fusion of these two scientific goals, but is just one example of how this approach could be used to create any similar platforms for science gateways or beyond.

The Science Library is built using technology that was developed in our earlier research (Controlled English, CE). This paper provides high-level details relating to the implementation illustrating how this approach can provide flexibility in extending the knowledge graph to capture complex relationships. The DAIS ITA Science Library at the time this paper was written contains 13,500 nodes, with approximately 105,000 links and 140,000 property values in the semantic knowledge graph. This large and complex graph represents almost 500 papers, over 400 authors, and 82 organizations; the corpus so far for the DAIS ITA program.

In Section 2, we describe the various ways to interact with the complex knowledge graph through visualizations, accompanied by the motivations behind them and in Section 3, we explain the CNL basis for the application, and the conceptual model schema that underpins the work. Section 3 also covers the Science Library evolution and implementation details, along with some future opportunities. Section 4 provides details of the administration mechanisms that we use to curate and validate the papers prior to publication, along with the motivations for each of the user interface designs for the activities we need to undertake. In Section 5, we discuss related work, for example, using CNL to drive Science Gateways, along with considerations relating to provenance and authenticity, and Section 6 concludes the paper. In this work, we do not attempt to offer any proof or validation of the potential popularity or scale of adoption that could be achieved with this approach.

2 \INTERACTION WITH THE KNOWLEDGE GRAPH

For many research groups or programs, a key output of the research activities is the academic publications created by the researchers. From these a rich social, structural, and topical network between authors and organizations emerges, and since the DAIS ITA is focused in particular on fostering a deep collaborative mind-set, it is important for us to understand and explore this aspect of our work. To capture and convey this body of published research and support this exploration, we have created the publicly available Science Library\(^1\) a user-centric, interactive portal. The idea for the Science Library originally came from the earlier NIS ITA research program, and was originally developed and deployed \(^2\) for that purpose, but has been reused and extended for this later DAIS ITA research program. Refer to Section 3.3 for details of the evolution of the Science Library and the timeline from the original NIS ITA implementation. The Science Library is a CNL driven research gateway, which allows the community to explore and query the publications and underlying networks through an open, web-based application built on a complex semantic knowledge graph. The data are represented through interactive visualizations along with the ability for users to query the model using a natural language query interface. This CNL-based approach models the data through concepts, properties, and relationships which are defined using CNL and are therefore both human readable and directly machine processable. This captures complex semantics in a simple format and enables non-technical

---

\(^1\)The Science Library for DAIS ITA (Distributed Analytics and Information Science International Technology Alliance) is publicly available at http://sl.dais-ita.org/science-library/.

\(^2\)The Science Library for NIS ITA (Network and Information Science International Technology Alliance) is publicly available at http://nis-ita.org/science-library/.
users to participate in the continuous improvement of the data model behind the Science Library application. This paper presents the features, implementation, and design considerations of the Science Library, the underlying CNL implementation, and the mechanisms for curation and administration.

One primary purpose of the Science Library is the exploration of publications from a given research program, such as the DAIS ITA. Figure 1 shows, on the left-hand side, a simple list of publications that can be sorted by various criteria including: recency, name, citation count. Additional information is available on this page, such as an overall summary of the publications for the program, broken down into publication type (journal, external conference, internal conference, technical report or patent). Each publication has a colored icon to indicate the program, and any paper variants are indicated by stacked icons. For example, it is often the case that an original journal or conference publication will be

**FIGURE 1**  List of publications (left) and conference paper details (right)
summarized or reworked into an internal conference publication for presentation to the DAIS ITA research community at the annual meeting of the alliance.

Throughout the Science Library application all entities (such as publications or authors) can be clicked on to take the user to a page that shows the details about that particular entity. On the right-hand side of Figure 1, we show the publication details page which provides a large thumbnail image of the paper and the ability to download all relevant information regarding that publication (e.g., the paper PDF, any presentation slides, poster, or other supporting material). We also display the author information, the paper title, and abstract and any venue information, including the location at which the paper was presented (for conference papers). All of this information is sourced from the semantic knowledge graph that underpins the system and is available for searching or exploration of the graph.

Figure 2 shows a narrative timeline for the publications of any author as well as a separate view of their co-authorship social network. The narrative visualization shows the development of research ideas and topics over time, along with co-authorship sub-groups. Users can read this visualization from left to right, with each node on the network being a particular publication at a point in time and the links between them showing co-authorship flows between the papers. This can be particularly useful to determine when new threads or topics form part-way through the research, and how they manifest in terms of publications over time.

Building on this is the more traditional co-authorship chart (also shown in Figure 2). The nodes represent authors, indicated by their initial (e.g., "DB" for Dave Braines), and are linked to other authors they have co-authored papers with. The thickness of the links indicates the frequency of that co-authorship, and this also drives the proximity of the nodes. In the interactive visualization, users can hover over particular nodes to have them highlighted, with extra information about the author in question appearing over the node. This is particularly useful for highlighting common co-authorship subgroups, which often indicate particular topics or threads of work that subsets of authors regularly publish. The colors of the nodes indicate whether an author is from academia, industry, or government.

Sometimes it is useful to see the specific publications for a pair of co-authors, that is, only those papers for which both authors are involved. This is shown on the left-hand side of Figure 3 and is represented as a stacked bar chart over time, with the bars representing the papers of a particular type. Hovering over any bar shows the individual papers for that period, enabling the user to click through to see their details. This is a very similar visualization to the one used for displaying the publications for any given organization and provides a simple visual mechanism for grouping publications over time. It should be noted that the visualizations chosen for each view are purely subjective based on our informal opinion as to what “works best” for the given situation and is not yet according to any scientific basis.

In Figure 3, we also show, on the right-hand side, the features of a conference at which multiple DAIS publications have been presented. This is mainly to showcase that we have the geo-spatial data available for analytics, rather than it being a particularly useful view. We show the venue of the publication(s) as a large yellow circle, and the authors (color coded according to their organization type) are shown in the geographic location associated with their organization, with the radius of the circles indicating the number of authors from each location.

Each page also has a simple search bar at the bottom of the screen. This will accept any search terms from the user and performs a reasonably complex semantic search into the semantic knowledge graph. If the terms that are searched for result in matches to specific named instances, then the result shown to the user is the relevant page on which this information can be located. For example the search term “Dave Braines
co-authors” will return a result which is the author page for Dave Braines in the co-author graph mode. In the event that multiple matches are made, or matches are made to which there is no predefined visualization type, then a simple list of clickable links (to papers, authors, organizations, topics, etc) are shown. The search is based on a combination of fuzzy keyword searching (to match title and abstract terms) as well as semantic graph searching, enabling more complex searches based on the graph relationships. This capability is underpinned by the “Hudson” natural language search APIs which provide a rich JSON response indicating which concepts, properties, or instances were matched in the semantic knowledge graph based on the query terms used. There is a lot of additional power that can be gained from using this approach which is not yet implemented into the Science Library. For example the ability to ask questions such as “list authors who have written papers with someone from industry on the topic of IoT before 2018.” Today, the interpreter API will match all of the terms in this question to the corresponding features in the semantic graph, but we have not yet developed a suitable “answerer” capability to take this interpretation and show the result to the user in a meaningful way.

In all of these visualizations, we have designed a simple user experience with consistent representations and actions, and have tried to showcase a variety of visual techniques to communicate certain aspects of the information in the underlying semantic knowledge graph. There is far more information available in the graph than we are making use of, making it possible in the future to create a much richer set of visualizations as requirements arise.

3 BUILDING THE SCIENCE LIBRARY

All of the visualizations and interaction features described in the preceding section are implemented using traditional web development technologies using HTML, CSS, and Javascript (client and server side), and they could be integrated with any suitable back-end database-like system. However, a key non-functional requirement of our solution is the ability to rapidly extend or transform the capabilities available to the end users as new opportunities are identified, or extensions to the captured data become possible. In order to achieve this, we need to build the system using dynamic components that are designed for such extensions from the start, and determine which parts of the overall system would require modifications, and to what extent, with each type of extension that could be added in the future. Each of these aspects, along with a description of the system and its evolution through various iterations from the original NIS ITA implementation are described in the sub-sections below.

3.1 Using controlled natural language

A key novelty in our approach comes in the technology used to implement the semantic knowledge graph, which drives the visualizations and allows users to easily contribute to, or extend, the model. To achieve this, we use a CNL technology named CE, which we defined and developed in previous research in order to provide agile knowledge representation and reasoning capabilities like those available in the Semantic Web stack. CE is a constrained form of English that is easy for non-technical human users to read, and it is possible for them to write it too (it is harder to write, but still less technical than traditional alternative machine formats such as RDF/OWL). CE is a language that is both human friendly and the directly machine processable language at the same time. There is no need to convert CE into some technical form (such as RDF or JSON) for the machine to process it. It is directly processed in its original readable English form. This is a key differentiating factor to our approach, and we believe is part of the reason why it is so agile as an ontology development and semantic graph building language.
there is a person named 'dave_braines' that has 'Dave' as forename and has 'Braines' as surname and has 'Dave Braines' as full name and has the organisation 'IBM UK' as default organisation and has 'http://...dave_braines.jpg' as profile picture.

the person 'dave_braines' has the person 'eunjin_lee' as co-author because the person 'dave_braines' wrote the document 'doc-3859' and the person 'eunjin_lee' wrote the document 'doc-3859'.

there is a document named 'doc-3859' that is an external conference paper and has The Science Library: A Controlled Natural Language driven Science Gateway as title and applies to the programme 'DAIS-ITA' and has the event 'IW5G 2019' as venue and has the project 'BPPI8_15' as project.

the published person 'dave_braines' wrote the document 'doc-3859'.

Figure 4 shows some CE examples taken from the DAIS IT A Science Library, illustrating the structure and readability of the sentences used to rapidly build up the concepts and relationships in the semantic knowledge graph.

We operationalize the CE language in the open source ce-store4 which is a simple Java-based web services component to allow developers to easily integrate with CE knowledge graphs using REST APIs that are familiar to them when using other technologies. The development environment is straightforward, and will be familiar to typical developers who can access the data using simple APIs. In our previous research, we have written extensively about the use of CE and ce-store in a variety of settings directly relevant to the usage in this paper. In earlier work,8 we define a conversational mechanism using this approach which enables users to converse with the underlying knowledge graph using natural language queries, and this is what led to the development of the "Hudson" APIs (as mentioned in Section 2) which underpin the Science Library search capability.

In addition to the core ce-store engine being available as open source, we have also open sourced the Science Library user interface5 and the CE conceptual models that underpin it.6 All the components and functionality presented in this work can be used freely by other repositories; the entire approach is open sourced and available for anyone else to adopt as separate, stand-alone Science Library instances. Full details and tutorials on how to set up a similar Science Library using CE can be found on the ce-store wiki page.7

Our insight into the value of rapidly defined conceptual models in real-time and the ability to therefore support the capture of complex semantic data came from our field trial work on various exercises.9 This evolved in various directions, one of which was the highly extensible capture of biographical and social information to aid human decision-making10 and this led us directly to the idea for building the Science Library to capture our own contextual and social network for our research publications.

3.2 Conceptual model for academic publications

The ability to rapidly define the underlying semantic conceptual models and then populate them with corresponding data enables us to build the Science Library application in a very short amount of time. We were able to develop simple models quickly, populate a small amount of data and then build some example visualizations in a very rapid manner and iterate improvements in a series of short sprints. The flexibility to modify and extend the concepts, properties or instances in the underlying graph gave us a very fluid environment in which we could rapidly explore new ideas and new conceptualizations before finalizing on those that worked. Figure 5 shows a small fragment of the underlying knowledge graph that is created in order to power the Science Library visualizations (on the left-hand side), as well as a section of the schema that defines the structure and semantics of the nodes and relationships in the graph (on the right-hand side).

3.2.1 Semantic knowledge graph

As mentioned earlier the Science Library for the DAIS IT A research program at the time of publication for this paper contains almost 500 papers from over 400 unique authors across 82 organizations. However the underlying knowledge graph is far more complex than this, consisting of 13,500 nodes, 105,000 links, and 140,000 property values which represent all of the information required to power the Science Library. On the left-hand side of Figure 5, a small section of this knowledge graph is shown, and it is from this large and complex graph that the power of the Science Library application arises. Much of the information in the graph is inferred using logical inference rules defined within the CE language, but all of this semantic and reasoning complexity is hidden from consumers of the graph (such as the code behind the pages that render the Science Library), meaning that standard web developers can easily work with the knowledge graph through simple APIs.

4 IT A CE store (ce-store), https://github.com/ce-store/ce-store.
5 Science Library user interface, available at https://github.com/ce-store/science-library.
6 Science Library data model for DAIS IT A, available at https://github.com/ce-store/sl-data-dais.
7 CE-store wiki and tutorials, available at https://github.com/ce-store/ce-store/wiki.
FIGURE 5 The underlying complex semantic knowledge graph (left) and the corresponding conceptual model or schema (right)

On the left-hand side of Figure 5, the green nodes represent topics, that is, the scientific or domain topics about which papers are written. The central dark blue node represents the "DAIS IT A" program (since the Science Library can support multiple programs within a single corpus of data), and the light blue nodes represent the various papers included within the Science Library. The orange nodes represent organizations and the pink nodes are the authors, who are linked to both organizations and papers. On the link between authors and papers, there is a small green node which indicates that within the graph there is a necessary intermediate object to capture specific information about that particular authorship, such as the position of the author in the author list, and the organization that is claimed for that particular publication (since some authors have multiple concurrent affiliations, or change affiliations over time).

While the knowledge graph is far more complex that the sample shown above (for example, it does not show structural information such as project affiliation, nor the venue for publications, nor citation details), it is a useful illustration of both the complexity and the value of the information stored within it. It also shows how this graph is a rich source of information for analytics of the publications over time and across various measures such as organizations, projects, topics, or venues. All of these analytic opportunities could be harnessed in the future by having analytic agents operate on the graph and the render their results as new nodes or links into the graph, thereby making them available for simple consumption by the pages of the Science Library or other means. This ability for both human and machine agents to contribute extensions to the knowledge graph using the same common language is inspired by the broad concept of blackboard architecture\textsuperscript{11} for multi-agent systems and we find this to be very valuable for creation of the Science Library knowledge graph.

3.2.2 Schema

Key concepts in the schema are: document (an academic publication), ordered author (an authorship of a paper), person (an individual who may author multiple papers), organization, event, and so on. Each of these concepts has one or more named relationships to other concepts, or textual values (such as title or abstract) which are collectively known as properties. These concepts and properties are the equivalent of a conceptual graph\textsuperscript{12} and the semantics are added through the definition of inheritance and the specification of logical inference rules. We use numerous rules in our conceptual model to infer additional information. Many times this is simply to more fully connect the graph to enable the development of the visualizations to be more straightforward, but in many cases they also add significant additional "logical value" to the model. For example we infer co-authorship from paper authorship: If two authors write the same paper, then they are inferred to be co-authors. In addition, some custom agents compute more complex values such as the local h-index\textsuperscript{8} for authors.

3.3 Evolving the system and implementation

Figure 6 shows the various phases of the implementation of the overall system, and how the information flows from our private content management system into the public Science Library application. When the original conference paper was written\textsuperscript{3} we were at phase 2 (labeled C:\slides\20130927\Datum\10.09.2013\fig4.pdf)

\textsuperscript{8}See https://en.wikipedia.org/wiki/H-index for a description of h-index.
"Improved" in Figure 6) and the discussion in that paper focused on the improvements achieved at that point. Here we are able to describe the latest improvements in the evolving system which are mainly focused around efficiencies in curation and administration and are covered in detail in Section 4. We also take a step back to the original implementation to show the full evolution of the system so far, and the insights gained at each phase.

At the bottom of Figure 6, the various major sources of information are shown. These are separated into these four categories because of the different roles that create the data, and the different ways in which they are created:

1. **Papers, authors, organizations, venues:**
   Created by administrators based on the PDF of the paper that is uploaded by the author. This requires validation of the paper by finding it in the relevant conference or journal proceedings, and manual creation of all the constituent parts of the publication.

2. **Citation statistics:**
   Obtained by administrators using Google Scholar.

3. **Paper PDF files:**
   Uploaded to the Drupal content management system by paper authors.

4. **Generated thumbnails:**
   Created by administrators for each paper PDF.

From the start we have designed a system to place minimum administrative burden on our consortium of paper authors, enabling them to simply provide a PDF of their publication along with details of where it has been published, and we as administrators manually compile all of the required information (authors, venue, organizations, etc) in order to publish this on their behalf. The publication of this data takes the form of appropriate nodes and links within the knowledge graph described earlier and in order to generate this without ambiguity we explicitly define each node and link, rather than assuming they can be reliably extracted from the unstructured information in the PDF documents.

Given these high-level goals the overall purpose of the Science Library has not changed since the original implementation, and the experience for the authors of the papers is similarly unaffected. All of the improvements in the two iterations have been on the efficiency, and therefore cost, of the administration of the system, and the generation of the Science Library semantic knowledge graph.

From an administration perspective, once the paper authors upload a PDF to the Drupal based content management system, the administrators manually create an instance of a Science Library publication and add the related attributes. A JPEG thumbnail of the paper is generated using the PDFBox tool and the citation statistics are manually pulled from Google Scholar and also added to the Science Library publication instance. Once all of these steps are complete the updated Science Library graph can be generated, enabling the end user interface to show the new papers, authors, topics, venues, organizations, and any other updates. This is achieved by using the ce-store to ingest the generated CE for the new publications, updating the existing instances and relationships.

This all sounds relatively straightforward and indeed is, but there are a number of important and time consuming factors that are overlooked in the simplistic functional description above. Many of these relate to assurance, information quality, and general curation of the full set of publications.
rather than the individual papers. It is these factors that have driven the evolution of the system shown in Figure 6 and where much of the cost and complexity actually lies. A couple of examples are listed below (but there are many such cases, often unique, and only discovered through the various assurance and cross-checking processes that we have put into place):

- **Confirmed publication:**
  Validation that the claimed publication is reported in the relevant conference proceedings or journal index. Sometimes papers are submitted and rejected, and the authors do not always remember to update the uploaded record in the Drupal system, so it is essential that we confirm each publication before inclusion. We capture the provenance for the paper in the form of a verification url property that shows the external evidence that the paper was published as claimed by the authors.

- **Pre-publication:**
  It is important that we do not pre-publish the work of any author. We therefore wait until after the conference or journal publication before publishing the paper within Science Library.

- **Missed publications:**
  Some authors forget to upload their publications to the Drupal system but these can be found using Google Scholar by searching on the program contract number in the acknowledgement.

- **In-scope publication:**
  All publications should include the contract number in the acknowledgement, but in rare cases this is missed by the authors. For these cases, we contact the authors to individually validate that the paper should be attributed to the program before we publish to Science Library.

### 3.3.1 Original implementation

In the original implementation (see Figure 6), we were at the end of the NIS IT A research program and decided to build this knowledge graph approach to the Science Library using the CNL technology that had been developed as part of that program. At this point we were using the Drupal content management system for storage of the paper PDFs, but the reporting of these papers (effectively the “index” that gives additional context such as time, place, project) was done using spreadsheets on a quarterly basis, the format of which changed over time, and the content of which varied in quality over time and across projects. The initial knowledge graph was created by manual analysis of 40 separate spreadsheets. These were created by 12 projects reporting progress every quarter for 10 years. The exercise was very time consuming but was a “one off” exercise for a program that was concluding, so at this stage there was no direct value in implementing an improved process for capturing information in the future.

During this original implementation, we also discovered that many publications had not been reported in the spreadsheets. In some cases, these were uploaded onto the Drupal content management system, so identifying these papers was fairly straightforward, although disambiguation between papers authored by the NIS IT A program and those uploaded for reference purposes was critically important as we cannot claim authorship of papers outside the scope of the program. We also found cases where papers had been published but not uploaded to the Drupal system or reported in the spreadsheets. These were identified by searching in Google Scholar on the program contract number that is included in the acknowledgement section in each published paper.

Having cleaned and manually curated the final set of publications for the NIS IT A program, we then generated the required knowledge graph in the CE language directly from the master spreadsheet using simple MS-Excel macros. We also manually retrieved paper citation data from Google Scholar and generated thumbnail images (using PDFBox) for each of the papers. This process is shown in the “Original” phase of Figure 6.

### 3.3.2 Improved implementation

Following the NIS IT A program, we wanted to apply the same technique to the DAIS IT A program, but we knew from the start that this would require substantial curation and administration and wanted to reduce the burden. We also knew in advance that authors might forget to report their publications and these could be found by searching for the contract number in the acknowledgement using Google Scholar.

To improve the administration and reduce the overhead, we moved away from spreadsheet based reporting and instead encouraged authors to report their publications via progress reports directly created on the Drupal system. We also created a new set of forms for administrators only within the Drupal system, enabling them to create all of the Science Library components (such as authors, venues, organizations, citation details) and link them directly to the publications uploaded by the authors. Limitations of the Drupal environment meant that these required one additional form per component, so while this is a substantial improvement on the previous spreadsheet-based system it does mean that papers with multiple
new authors from new organizations require many separate forms to be completed to model all of the new entities (papers with existing authors are much simpler, requiring the administrator to simply select the existing authors from a list).

This substantially improved the administration and reduced the overhead of doing so, but helped little with curation and validation which were still a largely manual process. For example, cross-checking publications found in Google Scholar using the contract number was still a tedious manual process using a temporary spreadsheet and handling fuzzy matches on paper names.

Another imperfection with this “Improved” system was the Science Library data now only being accessible in raw format via the online Drupal system. This was efficient for creation and for generation of the Science Library graph (using PHP macros within the Drupal environment) but was not as convenient as having all the data represented in a stand-alone data source such as a spreadsheet for archive purposes, offline working, or for easy sharing with other administrators or peer reviewers of the program. This process is shown in the “Improved” phase of Figure 6.

### 3.3.3 Current implementation

It was mainly the shortcomings described above that led us to develop the current implementation. The overhead of creating each Science Library component as a separate form within the Drupal system was a substantial overhead, and the need to have a network connection when doing so was another limiting factor. These two issues can be resolved through the use of a spreadsheet or similar structured data source that can be easily edited offline, but many additional issues around data quality and cumbersome cross-checking can be introduced by reverting to a spreadsheet format as per the “Original” implementation. We realized that a simple custom application with a well defined simple form to allow the administrator to rapidly create the required Science Library entities from the paper PDF would save substantial effort for paper creation, and would also provide the opportunity for some form of semi-automated extraction from the PDF where possible. We also realized that by saving the results of this process into a structured spreadsheet we could then have a portable and human readable format for the underlying data as well as an efficient way for creating and maintaining it.

The original PDF files as created by the authors on the Drupal content management system are unchanged and are the starting point for the process, and we made use of the emerging Cogni-Sketch\(^9\) environment to build the administration environment. This enables the administrators to use the Cogni-Sketch environment for efficient creation and navigation of the Science Library data ahead of publication, but also for this data to be persisted into a simple spreadsheet, achieving portability and potential offline editing (any changes to the spreadsheet are reflected in the Cogni-Sketch environment when it is reloaded). This also provides a much stronger basis for validation and curation, cross-checking against Google Scholar, and for preparing publications ahead of publishing but not pushing them to Science Library until after official publication externally. This new administration system also provides easy opportunities for embedding more advanced capabilities such as automated collection of citation data via APIs in the future. This process is shown in the “Current” phase of Figure 6.

Section 4 on administration and curation describes the current implementation in more detail, along with examples of semi-automated extraction from PDF files and automated suggestions for linkages into the existing knowledge graph. These small improvements to the administration have had a substantial effect on administrator productivity. By enabling the administrators to experience the underlying data in different views, we believe that the ability to identify issues and discrepancies is improved, leading to improved quality in the resulting knowledge graph as well as reduced time in maintaining it.

### 3.4 Future opportunities

In the original conference paper,\(^3\) we outlined a number of future plans, many of which have been implemented already and are reflected in the “Current” phase described previously. There are however a small number of remaining short-term opportunities for extensions that remain, as well as the broader generic opportunity for conceptual model extensions to support new fundamental capabilities or new data types.

#### 3.4.1 Platform support

As web development platforms and libraries have improved, it is possible to more efficiently represent and render some of the content of the Science Library, especially for mobile platforms or touch screen environments, neither of which are currently directly supported. The benefits here would include an improved experience for the end users as well as a cleaner presentation of the Science Library material to search engines, resulting

---

\(^9\)Cogni-Sketch is a tool being developed under the DAIS ITA program to explore human-agent teaming and efficient sharing of complex information as part of the research into future hybrid human-machine systems. It is currently available to limited users at https://cogni-sketch.org but will eventually be open sourced on github.
in improved rankings for search terms and therefore the potential for increased popularity of the Science Library and increased user visits. It is possible that some of the existing rendering and navigation choices could be revisited to better support users on touchscreen platforms, and for users accessing the site on a mobile phone platform or any other small-screen form factor a fundamentally different user experience, for example based around a single page design with simplified navigation\textsuperscript{15} rather than the more complex current solution that favors a larger screen environment.

3.4.2 Improved indexing of publications

Through use of the correct metadata within the Science Library pages, it is possible to improve the visibility of our publications within indexing sites such as Google Scholar and others. This is separate to indexing within search engines and is of particular relevance to some of our publications that are not published elsewhere and might otherwise be listed simply with institutional repository information. While the majority of the publications included in the Science Library are already well represented elsewhere in external sources (e.g., conference proceedings or journal indexes), we also run our own ‘Annual Fall Meeting’ (AFM) series which is an internal conference for the DAIS ITA research community. In most cases, the material presentations at these AFM events are papers that are previously published elsewhere at external peer reviewed venues, however the community does also create novel publications at these events too, either by substantial extension to previously published work, or outlining new emerging topics, or through adding additional domain relevance to a more scientific existing publication. In these cases, the work is novel to the AFM event and will not be well represented elsewhere and therefore is not indexed by sites such as Google scholar. To improve the visibility of these works\textsuperscript{16} and all of our publications more generally, we plan to use appropriate meta-data schema such as Highwire Press, ePrints or other tag systems. The information required to populate these tags is already available within our CE model described earlier, so making these available as meta-data tags within the generated pages is a simple but valuable extension.

3.4.3 Automated citation data

To better understand the impact of our publications, we measure the citations for each paper and use Google Scholar as the source for this information. There is currently no API available for Google Scholar so we collect this data manually for each paper on an approximately monthly basis. This is time-consuming but we have designed the new curation interface to make this as efficient as possible. In the future, it would be very simple to access this data via any API that may be made available, either in real-time (i.e., obtaining the citation data for each publication as it is requested by the users of the Science Library), or on a periodic basis as is currently done, but with the process being automated based on paper title search rather than being performed manually. This eventual automated implementation to a reputable citation API will be invisible to the end users of the site but will remove a substantial manual step required in the curation of the underlying dataset by the administrators and therefore will improve the overall efficiency of the solution.

4 ADMINISTRATION AND CURATION

Most of the improvements in the current phase of the Science Library are to support efficient and thorough administration and curation. This is a resource intensive operation and increases as the number of publications grow. As described in Section 3.3.2, the form-based solution using the Drupal based environment was an improvement on the original version but not as efficient as it could be. In this section, we describe in more detail the various “backend” improvements that have been created for administrators, with examples of their implementation.

A key principle in this work is that the administrators should have access to multiple interfaces to ensure they can choose the most efficient one for their task. While we have chosen interface designs based on our perception, and limited feedback from users, we can improve this variety of interfaces to match the cognitive task on a more scientific basis in the future.\textsuperscript{6} The tasks vary in intensity and purpose as well as frequency, and different formats and experiences are likely to be more useful depending on which task is being performed. We also want to maintain the overall agility of the CNL approach, meaning that we could evolve the Science Library system further in the future by extending the underlying model and corresponding user interface views, so any additional assets that we create for administration and curation must take this overall agility goal into account.

The fundamental user experience styles that we have made use of for this environment are:

- **Custom forms:** To provide improved efficiency and automate cross-checking (such as choosing valid authors, venues, and projects), we have created a small number of custom forms. These substantially reduce the manual burden of creating and linking items and reduce the chance of errors but they do create a fixed point in the system which is tied to the current conceptual model and may require updates if that model is advanced in the future.
• **Visual/graphical diagram:**
  The ability to render fragments of the graph as a relationship diagram is key to data exploration and understanding and can be used by the administrator whenever needed to get a deeper understanding of the data. They can also optionally use this view to add new node, links, properties, or attributes as needed. This provides some future-proofing against small conceptual model updates that can be supported using this technique instead of forcing updates to the custom forms.

• **Simple list:**
  These lists are filterable and can be editable if needed. This allows the administrator to limit the list to relevant items, and where necessary, quickly make edits to multiple items in rapid succession. These are ideal for low effort bulk updates where repetitive manual action is required, but not much analytic effort (e.g., citation data updates). This is flexible and can be applied to any content in the current conceptual model or future updates.

All of these are made available through the new Cogni-Sketch environment, which directly supports the visual/graphical diagram mode and simple filterable lists along with support for plugging in custom forms. The data in this environment are persisted by default into a specific JSON representation on the server, but an optional import/export to MS-Excel spreadsheet is also possible, providing the desired portability and offline editing capability mentioned earlier.

### 4.1 Custom form for creating publications

Creation of the Science Library documents requires multiple different data types and so a custom web form provides the ideal user interface experience to efficiently create all of the required nodes in the paper network. This task most benefits from additional automation to remove many of the manual steps that were present in the Drupal-based administration system.

A Science Library document is made up of different elements (Figure 8 shows a graphical representation of the component structure of a typical Science Library paper). In the previous solution, each element had to be manually and separately created in the Drupal environment. Each document for example requires a separate citation record, a link to each author (which need to be created if they do not already exist) and many more.

Automating aspects of the creation of the Science Library documents not only reduce the effort involved but allow some complexities of the system to be hidden from the administrator, and reduce potential errors being introduced, or items being missed. The custom form for creating a new publication is shown on the left-hand side of Figure 7 with various fields populated. The process can be entirely manual (with type-ahead look-up on the various fields to make finding related entities as fast as possible), or it can be semi-automated by processing the PDF file to identify any information that can be extracted. PDF files can be very different in their internal structure so it is not always possible, however in many cases the title, authors, affiliations, and abstract of the paper can be automatically extracted and populated into the form for administrator review and approval.

**FIGURE 7** Custom form for improved efficiency in creating new publications (left) and bulk-update form for information lists, for example, citation details (right)
To support this process, some of the underlying information in the knowledge graph has been augmented, for example, by adding pseudonyms for authors and organizations to capture the different ways that they are expressed in different publications.

The manual generation of the paper, poster, or presentation thumbnails from the underlying PDF files has also been automated so the administrator gets these created automatically whenever they upload the underlying file to the "new paper" custom form.

The end result is two-fold: A substantially improved process for creating new papers within the system (especially if automated extraction from the PDF is possible), as well as a simplified user interface using a single form where the data can be quickly and easily reviewed for accuracy rather than needing to navigate between multiple places.

In future work, should suitable APIs become available, we hope to automate the population of Google Scholar citation data both here and in the monthly bulk update step described in the next subsection.

4.2 Editable list for bulk updates

To keep the citation data current, we aim to review and update citations for the papers on a monthly basis. As the number of papers grows, it becomes increasingly costly to undertake this manual process. It is possible that in the future this could be automated through the use of suitable APIs but these are not yet available, so in the meantime we have made the process as efficient as possible to do manually.

This has been achieved by moving to a simple list format as shown on the right-hand side of Figure 7, enabling multiple records to be updated at once, removing the need to enter data common to all entries, such as the current date. Each entry is shown as a separate item in a simple list and the link to open the Google Scholar entry for each paper is located against each item, making it very easy to open the link, get the value, and enter it into the record. These can be saved individually or in batches as determined by the user.

4.3 Graphical diagram for user exploration

Our experience in curating a number of separate Science Library instances over a number of years is that administrator familiarity with the data is key in understanding typical publishing behaviors and therefore spotting anomalies. Figure 8 shows an example of a simple network visualization for this paper in the new Cogni-Sketch interface. This paper can be created in a number of ways, for example: by using the custom form to create...
the graph by analyzing the PDF and seeking user confirmation for all of the values, or by manually drawing the components using the palette on the left-hand side of the Cogni-Sketch interface, or by editing the underlying spreadsheet (or any combination of these).

The administrators can explore the publication network as well as other views such as showing all the publications for an author, including material that has not yet been pushed to public Science Library. They can also attach notes and comments, and share these with other administrator users, as well as paste in material from external sources such as email correspondence with the authors. This can be useful in validating any questions relating to the publications that are confirmed directly by the authors. All of this additional supporting information is available within the Cogni-Sketch environment, but is not published to public Science Library, nor included in the generated knowledge graph.

4.4 Integration with external sources

Having provided the administrators with these different views for creating, validating, and exploring the data we also provide a number of specific functional capabilities to support various import/export and integration activities:

- **Import/export to spreadsheet:**
  To provide a convenient and portable human-readable format for archive purposes and for offline editing if needed. Also very useful for any custom query or analytic processes that are well served by spreadsheet operations.

- **Google Scholar cross-reference:**
  The ability to import a spreadsheet of papers found in Google Scholar that match the search term of the program contract number (or any other search term for other instances of Science Library) and compare these against the publications defined in Cogni-Sketch. The result of the analysis is a list of any publications found in Google Scholar that are not defined in Cogni-Sketch.

- **Generate Science Library:**
  A simple function to generate the CE sentences that define the knowledge graph for the Science Library and publish these to make the information publicly available. This excludes any notes, comments or other "non-Science Library" items along with any publications held back from publication or whose publication date is in the future and entirely supersedes the previous generation mechanism directly from the Drupal-based administration interface using PHP macros.

By creating this fairly simple but highly extensible environment for collaborative administration of the Science Library publications, we have enabled a substantial improvement in the efficiency of the administrators, enabling them to publish updates more regularly and spend less time on onerous manual tasks such as seeking citation data and cross-validation with Google Scholar. The flexibility and extensibility of the Cogni-Sketch environment helps to retain the overall goal of agility in conceptual model extensions in the future, albeit with the caveat that the custom forms will require modifications if the conceptual model is extended (or the additional information required must be captured manually using the generic tabular or graphical interfaces).

5 RELATED WORK

There is limited literature around the use of CNL to drive Science Gateways and interactive visualizations, however methods to allow users to quickly and easily contribute to ontologies using CNL are known.17,18 There are also several techniques which aid users in exploring and querying ontologies using CE. These methods allow users to query a knowledge graph using CNL, such as Attempto Controlled English.19 Semantic Wikis capture knowledge about the data within pages and the relationship between pages, which allow semantic querying of the data. Some semantic wikis export the ontologies of the wiki as RDF or OWL, and parse it to a CNL to enable queries.20,21

6 CONCLUSION

In this paper, we have described the Science Library as an example of a Science Gateway focused on digital library and virtual community capabilities. We focus the examples on the DAIS ITA research program although different installations of the Science Library are in use across multiple projects and programs. A key novelty is the extensible basis on which the Science Library is built; creating a complex semantic knowledge graph using CNL. The data from this graph are presented to end users through a series of user interface visualizations that are described in this paper.
We focus in particular on the role of the administrators and the need for them to provide a high-quality curated and validated corpus of publications, and the techniques that we provide to enable them to do so. To minimize the overhead to the administrators, we have created a number of recent improvements and semi-automated key parts of the process, but ensured that the administrators retain overall control for the content of the corpus. The components required to create a Science Library are largely open sourced, with links and some details provided in this paper. At the core of the new administration and curation improvements is a new tool named Cogni-Sketch which is currently under development and will eventually be open sourced as well. We believe that the extensible basis on which the Science Library is built enables it to be taken forward in multiple directions by different groups seeking to address Science Gateway requirements. We believe that the core CNL basis for the conceptual model and semantic knowledge graph enables such changes and extensions to be made by less technical users than a traditional semantic graph system.

ACKNOWLEDGEMENTS

This research was sponsored by the U.S. Army Research Laboratory and the UK Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the UK Ministry of Defence or the UK Government. The U.S. and UK Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in the DAIS ITA Science Library at https://sl.dais-ita.org with a full explanation found in this paper.

ORCID

Dave Braines https://orcid.org/0000-0003-3296-0842

REFERENCES

1. Pham T, Cirincione G, Swami A, Pearson G, Williams C. Distributed analytics and information science. Paper presented at: Proceedings of the 2015 18th International Conference on Information Fusion (Fusion) IEEE, Washington, DC; 2015:245-252. https://ieeexplore.ieee.org/document/7266569.
2. Preece A, Sieck WR. The international technology alliance in network and information sciences. IEEE Intell Syst. 2007;22(5):18-19.
3. Lee E, Braines D. The science library: a controlled natural language driven science gateway. Paper presented at: Proceedings of the 11th International Workshop on Science Gateways (IWSG), Ljubljana, Slovenia; 2019.
4. de Mel G, Braines D, Thomas A, Pham T, Dron W. Cognitively mediated research discovery: a context-aware rich visualized knowledge graph co-created by humans and machines using a common language. Paper presented at: Proceedings of the Workshop on Hybrid Human-Machine Computing (HHMC); 2017. Guildford, UK.
5. Kuhn T. A survey and classification of controlled natural languages. Comput Linguist. 2014;40(1):121-170.
6. Polson PG, Lewis C, Riemann J, Wharton C. Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. Int J Man-Mach Stud. 1992;36(5):741-773.
7. Mott D. Summary of ITA controlled English. ITA Technical Paper; 2010. http://nis-ita.org/science-library/paper/doc-1411a. Visited on May 11, 2020.
8. Preece A, Braines D, Pizzocaro D, Parizas C. Human-machine conversations to support multi-agency missions. ACM SIGMOBILE Mob Comput Commun Rev. 2014;18(1):75-84.
9. Braines D, Mott D, Laws S, de Mel G, Pham T. Controlled English to facilitate human/machine analytical processing, Proc. SPIE 8758, Next-Generation Analyst 2013:875808. https://dx.doi.org/10.1117/12.2015907.
10. Braines D, Ibbotson J, Shaw D, Preece A. Building a “living database” for human-machine intelligence analysis. Paper presented at: Proceedings of the Information Fusion (Fusion), 2015 18th International Conference on IEEE, Washington, DC; 2015:1977–1984. https://ieeexplore.ieee.org/document/7266797.
11. Weiss M, Stetter F. A hierarchical blackboard architecture for distributed AI systems. Paper presented at: Proceedings of the 4th International Conference on Software Engineering and Knowledge Engineering, Capri, Italy; 1992:349–355; IEEE.
12. Mognier ML, Chein M. Conceptual graphs: fundamental notions. Revue d’Intelligence Artificielle. 1992;6(4):365-406.
13. Lehmkühler A. Apache PDFBox-working with pdfs for Dummies. 2010.
14.Drupal C DORTMUND FDS. About Drupal, Drupal Camp; Orlando, FL: 2014.
15. Yu N, Kong J. User experience with web browsing on small screens: experimental investigations of mobile-page interface design and homepage design for news websites. Inf Sci. 2016;330:427-443.
16. Arlitsch K, O’Brien PS. Invisible institutional repositories: addressing the low indexing ratios of IRs in Google Scholar. Libr Hi Tech. 2012;30(1):60-81.
17. Tábanl V, Polajnar T, Cunningham H, Bontcheva K. User-friendly ontology authoring using a controlled language. LREC. European Language Resources Association (ELRA); 2006:35-40.
18. Hart G, Johnson M, Dolbear C. Rabbit: developing a control natural language for authoring ontologies. Paper presented at: Proceedings of the European Semantic Web Conference, Springer, New York, NY; 2008:348-360. https://link.springer.com/chapter/10.1007/978-3-540-68234-9_27.
19. Bernstein A, Kaufmann E, Göhring A, Kiefer C. Querying ontologies: a controlled English interface for end-users. Paper presented at: Proceedings of the International Semantic Web Conference; 2005:112-126; Springer, New York, NY.
20. Schaffert S. IkeWiki: a semantic wiki for collaborative knowledge management. Paper presented at: Proceedings of the 15th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE’06), Manchester, UK; 2006:388-396; IEEE. https://ieeexplore.ieee.org/document/4092241.

21. Buffa M, Gandon F, Ereteo G, Sander P, Faron C. SweetWiki: a semantic wiki. J Web Semant. 2008;6(1):84-97.

How to cite this article: Braines D, Stockdill-Mander J, Lee E. The Science Library: Curation and visualization of a science gateway repository. Concurrency Computat Pract Exper. 2021;33:e6100. https://doi.org/10.1002/cpe.6100