Linked Research: An Approach for Digitising Scholarly Communication

Sarven Capadisli[1]
[1]Enterprise Information Systems Department, University of Bonn, Bonn, Germany
info@csarven.ca

Amy Guy[2]
[2]School of Informatics, University of Edinburgh, Edinburgh, UK
amy@rhiaro.co.uk

Christoph Lange[3]
[4]Enterprise Information Systems Department, University of Bonn, Bonn, Germany
langec@cs.uni-bonn.de

Sören Auer[4]
[4]Enterprise Information Systems Department, University of Bonn, Bonn, Germany
auer@cs.uni-bonn.de

Nicola Greco[5]
[5]Decentralized Information Group, CSAIL, MIT, Cambridge, US
ngreco@mit.edu

DOCUMENT ID
http://csarven.ca/linked-research-scholarly-communication

IN REPLY TO
ACM Hypertext 2016 Call for Contributions

MODIFIED
2016-02-17

NOTIFICATIONS INBOX
inbox/

PUBLISHED
2015-12-19

ABSTRACT
The future of scholarly communication involves research results, analysis and data all being produced, published, verified and reused interactively on the Web, with 'papers' linking to and from each other at a granular level. The academic process of peer review is increasingly becoming open, transparent and decentralised. More broadly, the mechanism for global knowledge sharing is becoming an ongoing conversation between experts, policy makers, implementers, and the general public. This vision is radical, and getting there requires understanding of, and change in, a number of interrelated areas. In this article we break down the problem space and define requirements for advancement towards a Web-based ecosystem for scholarly communication: *Linked Research*. We discuss our strategy for tackling each of these areas. This includes how we can build on and combine existing well-known technologies and practices for digital publishing, social interactions, decentralised data storage, and semantic data enrichment. We provide an initial assessment of our proposed strategy with an example implementation of tooling which sets out to meet the requirements.

Keywords
Human-computer interaction, Linked Data, Semantic publishing, Social machine, Social web, Web science

1. INTRODUCTION
"You never change things by fighting the existing reality. To change something, build a new model that makes the existing model obsolete" – attributed to *Buckminster Fuller*.

One of the most widely debated questions in the scientific community is the impact of digitisation on scholarly communication and knowledge exchange; the Internet and the Web have radically changed the processes of scholarship [1]. In this article, we propose that digitising, and indeed *Webizing* [2], scholarly communication can provide greater coherency between and within academic fields, as well as make knowledge more accessible to researchers and citizens alike. Many forms of scholarship are enabled by these technology changes, be it the ability to perform social science using social media traces, or large scale data processing for physics. Despite efforts through Open Access movements and the increasing availability of online publications, the formal communications mechanisms used by scholars do little to take full advantage of either the capabilities of online media, or the cultural shift towards sharing and public commentary on social networks [3, 4.

On examination, it becomes clear that there are many interrelated problems to be tackled before it will be possible to fully realise the possibilities offered by modern Web technologies in the academic space. Many of these problems are relevant in a number of domains outside of scholarly publishing, and they are each to different degrees being considered in existing or ongoing work, in both academia and industry. Our first contribution is an overview of different issues, briefly summarising the extent to which they are addressed in current work, and how they relate to and depend upon one another in the context of scholarly communication. We attempt to articulate a complete picture of the problem space of scholarly communication by defining *requirements* in response to the problems described in existing literature, or observed 'in the field'. These requirements fall into three broad categories:

- Publishing and ownership of documents and data.

- Discovery and reuse of knowledge and data.
- User experience and tooling.

From these requirements, we derive user stories, and explain how these may be repurposed as an *acid test* or criteria against which to evaluate proposed solutions. Our strategy for addressing the problems identified involves building upon existing technologies, standards and best practices, each of which we discuss in detail in section 4.

Finally, we describe an implementation of the Linked Research concept for authoring, publishing and annotating research articles in a decentralised manner, based on native Web technologies. We demonstrate how multi-modal content such as video, audio, code examples and runnable experiments can be embedded into research publications, and showcase how different views of the content can be rendered for different devices (e.g. screen, print, mobile) or audiences (e.g. slide shows). A particular strength of our approach is the integration and linking of data, which can automatically update tables and diagrams when the underlying data changes. All content can be annotated and represented using semantic knowledge representation formalisms to facilitate better search, exploration, retrieval and linking of concepts and ideas. Further, our implementation takes advantage of common social media practices to enable sharing and commentary around scholarly work in both ongoing dialogue between interested parties, and more formal conference and workshop settings. We build upon emerging standards in personal data stores, annotations and social notifications to allow our implementation to be completely decentralised, reliant on no central system or hub.

2. REQUIREMENTS

The Linked Research requirements provide a framework for thinking about how each of the various parts of the problem space of scholarly communication are connected to each other. By establishing a framework, Linked Research can be seen as both a benchmarking exercise as well as a technique to connect various technologies and approaches. The topic areas are described below, and we have identified several interdependencies between them, whereby aspects of one problem are much more easily solved alongside of or on the back of another.

Socially, Linked Research is embedded within the Open Science movement [5]. However, it explicitly calls out the role of technical choices in meeting open science goals.

2.1 Publishing And Ownership

Open Access. To better enable scientific progress, research work, from raw results to completed analysis, should be widely available and accessible. Maximising reach in this way increases the likelihood of further scientific progress and new discoveries [6]. Ideally this is achieved by way of free-of-charge access and reuse of all material, including code, media, source text, data and metadata. This requirement is borrowed from the well-known *Open Access* movement, however this still often contains restrictions (e.g., Gold Access / Author Process Charges) imposed by third parties such as publishers, inhibiting the distribution of knowledge to any demographic, regardless of wealth, social status, or geographical location of either the researcher or the consumer.

Support for decentralised authoring and publishing. Research is commonly published through centralised services today. Publishers of journals or conference proceedings act as gatekeepers to work, which may only be available behind a paywall, and is required to meet criteria specific to the publishing house before it is published. Existing repository services (such as arXiv or Zenodo) which allow researchers to upload their own work directly to a database require proprietary processing (e.g. to extract metadata), account creation, and may distribute work only to other users of that service. If such a service shuts down, links to the work and any additional value extracted from it by the service may be lost if there are no other copies. In both cases, control over access and distribution of the work is taken out of the hands of the original author, and the author may additionally be required to give up their rights to distribute their work by other means. We propose that it must be possible for authors to opt out of such information or functional silos [7] and publish work on their own terms, in a storage location controlled by themselves or someone they trust, without being excluded from the wider community.

Access control and attribution. In many fields, researchers are discouraged from publishing unfinished, inconclusive or ongoing work, or publishing outside of 'official' venues by the risk that their work will be plagiarised or reused by another without permission or due credit. To facilitate open scientific discourse, technologies for publishing must afford a trust that sharing work will not disadvantage the author. Thus the author must have means to unambiguously attach their identifying information to their own work, and selectively grant and revoke access to others during the research process. Presently there are many services offering identifiers to creators of scholarly work: the *Library of Congress Linked Data Service* for personal and family names and corporate bodies, *Virtual International Authority File* (VIAF), *International Standard Name Identifier* (ISNI) and the Getty's *Union List of Artist Names* (ULAN), *ResearchGate*, *Google Scholar*, *Scopus Author Identifier*, *ResearcherID*, and *Open Researcher and Contributor ID* (ORCID). These systems are minimally, if at all, interoperable without human intervention, so creating multiple profiles is common practice. Some of these systems support – not via open standards but by individual solutions – certain others, e.g., ORCID can import data from ResearcherID. It is part of our vision to empower authors to choose an identity provider they trust, and use a single preferred identifier across their publications.

Provenance and accountability of information. Trust and confidence in data and results can be fostered by enabling reproducibility of experiments, and demonstrating a coherent explanation of the derivation of conclusions, and the agents responsible; however this tends to be difficult under current practices. Capturing scientific processes, workflows and data origins in a machine-readable way will further improve the utility of researchers' work.

Persistence and long term preservation. Research results must be made available in such a way that access to it is reliable and consistent over time, and irrespective of place. Broken links, missing data, authors who move institution, organisations which close or merge or rename, and regimes which censor can all hamper access to and reuse of knowledge. A Linked Research ecosystem strives for persistent, long-term access to information as a technical and social norm.

Sharing and social interactions. Authors largely rely on third parties, including publishers and the media, to distribute and raise

awareness their work. However, increasing numbers of academics are making use of social media to this end, and sharing on the Web has become a normal part of daily life for many [8]. Linked Research will capitalise on this trend.

Commentary and feedback. The academic peer-review process has been used to judge whether work makes contributions to the field. Studies are beginning to show the unreliability and elitist nature of this process [9, 10] and thus we call for transparency and broader participation in research discourse. It should be possible and expected for anyone to review academic work, and for this review to be attributed and recognised as a contribution of the reviewer in its own right. Moreover, research has shown that transparency in the peer-review process may be an indicator of the quality of peer-review [11].

2.2 Discovery And Reuse

Human and machine-readability. Scholarly content, mostly published in PDF form, is currently not machine-readable to the extent that automatic post-processing is possible. Desirable modes of post-processing include sorting, aggregation and more accessible display of information, to improve the ease of use by humans. Many aggregation and indexing systems require manual input of metadata (or careful scraping by proprietary software), and authors undertake this labour, often repeatedly across different systems, for the pay-off of improving findability of their work. This labour can be avoided by generating metadata from article content and publishing it according to a standard syntax and semantics.

Integration of rich semantics. Within and around the prose, scholarly communication comprises rich semantic structures:

- the structure of the content, e.g. in sections, examples, definitions, theorems, notations;
- the argumentation, e.g. premises, assumptions, deductions, conclusions;
- facts and relationships between entities expressed in sentences;
- all kinds of data (examples, experiment results).

Notably, the *Joint Declaration of Data Citation Principles* [12], endorsed by over 100 organisations, explicitly calls for data to be treated the same as papers in the scholarly ecosystem, and for machine readability of the data. Similarly, the *FAIR Data principles* for findable, accessible, interoperable and reusable data encourage the semantically interoperable reuse of data.

Making this level of semantic structure available to machines for querying (and ultimately automated reasoning) facilitates development of intelligent support services for scholars, such as recommendation, search and even comparison for related work.

Globally unique identification of entities. A research document contains references to *things*, from abstract concepts to real-world objects; anything which may be referenced from a publication (people, methods, diseases, phenotypes, organisations, and so on). For the scholar it must be possible to unambiguously indicate the subjects of discussion by creating their own identifiers for these entities or by reusing existing identifiers coined by others. This will facilitate associates between content referring to the same entities, which currently is a challenging and error-prone task involving natural language processing techniques (e.g. named entity recognition and disambiguation).

The progressive adoption of the DOI infrastructure for citation of artefacts other than articles is a strong step towards this area of Linked Research. Indeed, we are now beginning to see a linked graph of information around publications based on these DOI focused infrastructures, e.g., *Crossref & the Art of Cartography: an Open Map for Scholarly Communications*.

Integration of data. In many cases some form of structured data is a direct subject of scholarly communication. In social sciences, for example, statistical data plays a key role as ground truth for validating or falsifying theories. Similarly, in engineering and natural sciences, measurement or observation data is of crucial importance. In the life sciences, statistics, anonymised patient data as well as taxonomic and ontological data can be central. Ideally, such data is directly integrated into the digital representation of a scholarly article in both human- and machine-readable forms. This facilitates both live updates to the article (including charts and figures) as new data emerges with no extra effort to the authors, and the ability for closer exploration of the data for better understanding by consumers.

2.3 User Experience And Tooling

Integrated authoring and publication workflow. Researchers should be enabled as far as possible to have a seamless experience when making their work available to their community and to the general public. This spans from basic authoring of text content according to the desired layout, to embedding data and interactive elements, to managing citations, footnotes and metadata, and finally to publishing the article in the preferred location.

Feedback and interactions. When others interact with published work, for example to cite, comment, review, share, recommend or annotate, the original authors should be notified and able to view and react to any new content produced in response to their work. Conversations and commentary may be stored across different systems, however there must be a mechanism for aggregating and displaying scientific discourse around the subject resources to benefit authors and consumers alike.

Support for different views. Content is consumed in a wide variety of formats online and offline, on devices of different sizes and capabilities, and with different interaction techniques (touch screens, speech input/output). Scholarly communication should exploit these different media, presentation and interaction techniques to allow scholarly content to be consumed in different situations [13]. The content should adapt to the available presentation capabilities, and alter its appearance according to some combination of the author's and consumer's preferences.

Adaptation to audiences. Depending on the audience, presentation of scholarly communication should happen on different levels of granularity. For experts in the field, explanations can be shorter, examples are not required and only few illustrations are needed. For newcomers on the other hand, more detailed explanations, illustrations and examples are helpful. Ideally, scholarly content can adapt to the audience. This should happen from a single source of content to minimise the burden on the author, where content elements are marked in such a way the the adaptation to the audience can be performed largely automatically.

Integration of interactive content. Where possible, scholarly con-

tent should provide dynamic and interactive content. Examples include:

- executable software source code snippets, which demonstrate algorithms, queries, workflows etc. ([14] calls for these to be considered as a critical research output).
- interactive data visualisations with different diagram types, which allow users to zoom in, filter etc.
- small games or self assessment tests, which allow readers to interact with the content in a playful way or to test their comprehension of the knowledge
- widgets or applications, which provide interactive domain specific interactive content (e.g. exploring a large phenotype taxonomy)

Support for multimedia. Multimedia content such as videos, audio, and 3-d simulations can dramatically improve the comprehension of scholarly content compared to the traditional static 2-d illustrations currently found in papers. In particular, fields dealing a lot with multimedia data like engineering, arts, audio and video analysis, and medicine, would benefit.

Impact metrics and reward systems. CL: In this context we could more generally speak about measuring *quality*.

2.4 User Stories

We describe an *acid test* so that systems can be evaluated against the requirements of Linked Research previously outlined. The evaluation is intended to verify the openness, accessibility, decentralisation, interoperability of approaches in scholarly communication, and takes the form of a series of user stories which cover the full spectrum of scholarly communication and which must be feasible to carry out with the proposed system. This test does not mandate a specific technology, therefore the challenge can be met by different solutions. It is intended to test the design philosophies so that different approaches may be closer to passing the *independent invention* test.

This acid test is extended from *Enabling Accessible Knowledge Acid Test* [15] to encompass more aspects of the academic workflow, beyond authoring and publishing research articles. The acid test constitutes *assumptions* about proposed solutions and *challenges* that proposed solutions must meet one or more of.

ASSUMPTIONS

- All interactions conform with open standards, with 1) no dependency on proprietary APIs, protocols, or formats, and 2) no commercial dependency or priori relationship between the groups using the workflows and tools involved.
- Any mechanisms are available through at least two different interoperable tool stacks.
- Information and interactions are available for free and open access with suitable licensing and attribution for retrieval and reuse.
- Information is both human and machine-readable.
- All interactions are possible without prior out-of-band knowledge of the user's environment or configuration.

CHALLENGES

1. Alexander makes his article available on the Web with the research objects available at fine granularity, e.g., variables of a hypothesis.

2. Beverly refers to and discusses Alexander's research objects from her own research article, e.g., an argument against a methodological step.

3. Carol annotates Beverly's argument on Alexander's work by suggesting that it was misinterpreted, and stores the note publicly at her own personal content store.

4. Darmok and Eve write reviews for Beverly's article and store them at their preferred locations respectively; Beverly is notified about the reviews.

5. Frank publicly announces a call for contributions for an academic conference. He specifies the scope of the call and desired qualities of the submissions.

6. Guinan notices Frank's call for contributions to be suitable for her research article, and submits a link to her work. In her article, Guinan states that it was a reply to the call.

7. Frank assigns Herman, Inigo, and Jean-Luc to peer-review Guinan's work. Herman and Inigo both write their reviews so that only Frank (and Guinan?) may read them; meanwhile Jean-Luc makes his review public.

8. Keiko is assisting Frank, and compiles a list of articles which meet the requirements and standards of Frank's call for contributions based on reviews and public feedback. He sends alerts to various institutions that the proceedings are ready to be archived.

9. Liz manages scholarly archiving at her institution; she retrieves and catalogues the articles which were mentioned for library indexing.

10. Marshall is a PhD student; he uses the interactive components in Beverly's article by changing the parameters and rerunning an experiment and decides to expand on this for his own work.

11. Nelson reads Alexander's article from his handheld device, and prints a single page summary.

12. Ophelia notices the new index of articles in response to Frank's call and selects the ones she is interested in for her personal collection of potential references.

13. Paris sees the peer-reviews about Guinan's work and proceeds to discover further information about the qualifications and experience of the reviewers.

14. At Frank's conference, Q listens to the presentations from the authors and makes personal notes about the work; he occasionally makes his observations visible to the the authors or rest of the conference audience.

3. CONCEPTS

Having defined requirements for the problem space of scholarly communication as a whole based on observations and existing work, we now consider how each topic area fits into the overall scholarly communication workflow, and indicate which of the requirements are pertinent for each stage of the process of scholarly communication. We describe expectations for each stage of the process, and propose technical solutions and methods aiming at meeting the related requirements. The Linked Research concepts and how they relate to one another is shown in Figure 1, and their relation to the requirements, technologies and stakeholders are visualised in Figure 2.

Our primary approach is to reuse existing technologies, techniques and specifications, combined in novel ways to meet the needs we

have identified. We focus on the Web as a publishing platform for distribution and consumption and content and data in a decentralised way.

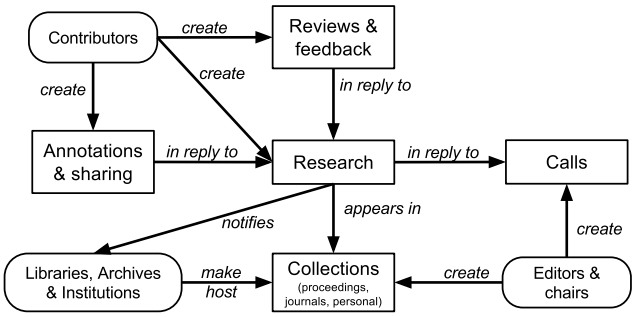

Figure 1: Linked Research ecosystem.

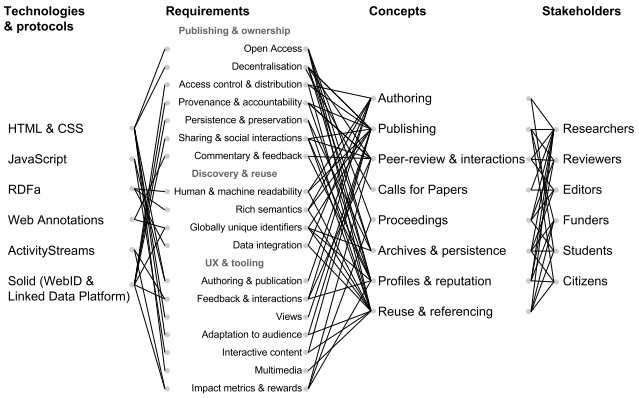

Figure 2: Links between technologies, requirements, concepts and stakeholders.

3.1 Research Authoring

Frontend Web technologies – e.g., HTML, CSS and JavaScript – can be purposed towards authoring tools that are straightforward to use for authors and that require no proprietary software or special configurations; all that is needed is a Web browser. Such tools can be used to edit local files when the author is in *offline mode* or without an internet connection, but can support writing directly to a personal data store or server when the author is online to facilitate collaboration. Authors can easily go beyond textual content, and embed data, interactive or multimedia elements into their work with native Web components. This is a well-established practice in the social Web, e.g., blogging. Authoring tools built on the Web can be extended or built upon with little overhead to meet specific needs of researchers.

Semantic Web technologies have been widely embraced by the librarian community for organising traditional scholarly content as well as exposing the rich terminological systems underlying such content [16]. For example, the BIBFRAME initiative aims to ease the transition from MARC metadata records to web enabled metadata systems, and both CrossRef and DataCite DOIs support Content Negotiation allowing multiple representations including machine readable representations to be returned.

We propose to use RDFa embedded in HTML together with established vocabularies to add fine-grained semantics to prose content, including metadata (e.g. authors, keywords), document structure (e.g. abstract, conclusion), components and concepts being described (e.g. methodological steps, hypotheses, specific instances of things). A browser-based authoring tool should both automatically generate RDFa where possible, and provide a user-friendly interface to allow the author to add semantics themselves.

3.2 Research Publishing

The Linked Research initiative encourages researchers to work transparently, publish early and often and without asking for permission. The Web provides the perfect platform for this. Ideally authors should acquire their own domain name and Web hosting; we acknowledge that this is not always possible or practical, though many institutions provide basic Web hosting and a URL space for all of their staff and students. Hence, employing DOI, PURL, w3id, or alike are possible ways to have some level of permanence for the information.

Publishing work online in static HTML allows wide accessibility for consumers and for added value services. Using CSS and JavaScript makes it possible to refine, adapt or even provide alternative views on a research document according to the consumer's needs or viewing context.

Distributing work with URLs allows sharing equally between academics and specialists, and interested citizens. Linking between articles enhances discoverability, and coupling this with social notifications *ActivityStreams 2.0* is an appropriate emerging W3C standard which would enable interoperability with other decentralised social systems) allows authors to track citations, the reach of their work, and receive and display responses and feedback.

It is important to include licensing information with published work in order to promote fair use and remixing for content creators. We do not presume to dictate the most appropriate licenses for academic work to be released under, however it is worth emphasising that establishing common practices around open licenses such as those from the Creative Commons family is important to encourage a culture of openness and reuse, which, as we previously discussed, is important for a healthy scientific progress.

3.2.1 Calls For Contributions

Calls for contributions to journals and conferences are usually published online, and an increasing number of venues are accepting HTML submissions. A step further to encourage ownership and control over research work is to accept submission by URL, whereby the authors submit notifications of a ‘reply’ to a call.

While most calls for contributions are available in plain and simple HTML, they are typically not semantically enriched. Making their semantics explicit and then publishing them in compliance with the Linked Data design principles would improve their interlinkability with other knowledge graphs such as research articles or social

feedback. Composing the call with additional structured data, for instance with concepts from controlled taxonomies (e.g., SKOS, DBpedia), topical requirements, and a list of contributors for peer-reviewing, enables better mining of such data. It can be purposed towards discovery calls suitable for a paper draft (i.e. *venue recommendation*), relating one call with similar calls, as well as watching the evolution of topical trends in research communities [17].

3.2.2 Peer-reviews And Interactions

In a Linked Research community, reviewers are able to see replies to a call they have, or have *been*, subscribed to, and in turn leave their reviews as replies to the submitted articles. Rather than centralistic submission and review management systems such as *EasyChair*, we employ workflows for 'sharing' that people are nowadays used to from social media. As peer-reviews are valuable contributions to the academic ecosystem, we finally encourage reviewers to *publish* their reviews in a Web space that they control (or trust), for the benefit of the community. With the optional adoption of attributed peer-reviews, responses can be rigorous and objective as they can be, while retaining transparency and identity of the reviewers. These open reviews are also subject for the community to discuss and build on.

3.2.3 Proceedings

Editors and committees who have issued calls for contributions are able to take into account reviews and discussion on work they have been notified about, and select the articles to aggregate and formally endorse them as part of their journal or event proceedings. The same Web-based tooling which allowed authoring of articles, calls and reviews and notifications of responses in the first place can be purposed to automatically assemble and re-publish the chosen work as a collection, which can in turn be shared and responded to.

3.3 Persistence And Preservation

Ensuring the persistence of scientific results through time is of fundamental importance: readers should be able to access articles published in the past, authors should be able to point to articles and the statements within that will continue to be accessible in the future. As of today, archiving and preserving scientific documents are commonly considered the responsibilities of publishers. In the context of Linked Research, where decentralised authoring and publishing is promoted, responsibilities are to be re-evaluated: authors can publish their results online at a service of their choice, such as their own website. However, persisting academic publishing on long-term on the web has two main challenges which may be difficult for individuals to solve when hosting content themselves: 1) availability on the Internet (keeping the document/data server available, maintaining it, and defending it from attacks), and 2) presence on the Web; namely, maintaining a URL that points to the document/data and that does not change. Long-term obstacles to this include URL expiration, revocation (in case of someone taking legal action) and censorship [18].

We already see many efforts at centralised archives - *Internet Archive*, *Archive-It*, *arXiv*, *CLOCKSS*, *Zenodo* - however these are each a single point of failure for long-term persistence. In the case of self-archiving, the author is responsible for maintaining their service and domain name. This strategy clearly has non-trivial disadvantages: technical expertise is required, there are maintenance costs, and content may disappear if the author loses the ability to keep this up. In the case of third-party archival, although the responsibilities are delegated a service, this may cease to exist long term, or may act against the interest of the user, for example altering the content.

An architecture in which documents are replicated on multiple servers can work towards solving these problems. If the URL is not reachable, the reader should be able to get a copy the content from elsewhere. "Lots of Copies Keep Stuff Safe" is the idea behind LOCKSS, a project that aims to create a network of libraries and publishers that replicate scientific articles amongst the others [19], based on this idea there is other work that uses existing distributed networks such as BitTorrent [20].

When content is distributed across a decentralised network, documents must have pointers that the network agrees upon. This can be simply done by naming documents via their hashes, as in the InterPlanetary File System (IPFS). This also allows proof that the the file hasn't been altered by nodes in the network, since verification would mean re-hashing the file. Recent works show how we can use *Trusty URIs* to employ hashing techniques whilst remaining compatible on the Web [21].

Ultimately in a Linked Research ecosystem, the role of content creators is also to distribute the content, and a role of public institutions in supporting preservation is to provide reliable hosting, addressing and backup in a decentralised manner. We envision a two-way system. Organisations such as libraries and universities can actively create collections by selecting, sorting and storing published articles. Additionally, authors can submit work themselves to archives, for either consideration by curators or automatic inclusion for backup. Currently in order to do this scholars may submit their work manually into different systems, which can result in inconsistent or infrequent updates. To improve on this, the same open Web protocols as described for calls and proceedings can be used to automatically notify chosen archives whenever a new piece of work or dataset is ready.

3.4 Profiles

Just as authors can choose domains they trust to identify their work, they can also create URIs to identify themselves. By connecting this identifier to all of their publications, results and impact of the work they are involved with can be traced directly back to them. Researchers can build a profile which automatically includes their published work and their feedback on the work of others.

Currently, many academics choose to create professional profiles on centralised systems such as *ImpactStory*, *Google Scholar*, *ResearchGate*, *Zenodo* or *Academia.edu*. Such systems are poorly integrated with each other, and users frequently must enter their details repeatedly or choose one system and miss out on potentially connecting with members of others. Similarly from within one system, users cannot see the activities of users in other systems, even though many of the types of social activities afforded by the systems (reading or recommending a paper for example) are the same.

4. REALISATION

In this section we describe our own implementations, which aim at fulfilling some of the requirements stated previously. We pay particular attention to rich, semantic authoring, and decentralised publishing of both content, data and feedback. We describe the technologies we used to meet each requirement, and how we combined cer-

tain technologies and approaches in novel ways. We conclude with a brief discussion of requirements yet to be met and how we plan to go about tackling these.

4.1 Authoring And Publishing

dokieli [22] is an open source, progressively enhanced client application for decentralised authoring, with specialisations for writing academic documents. The editor is built on open Web standards and the documents are compliant with Linked Data best practices, allowing: decentralised storage and data ownership; fine-grained semantic structure through HTML+RDFa for prose content, and other RDF formats for data embedding; direct in-browser editing from a W3C *Linked Data Platform* (LDP) based personal data store; social interactions with documents (such as annotations and replies), and notifications thereof.

dokieli is compliant with *Solid*, a set of protocols and conventions based on the LDP recommendation. This incorporates authentication, access control and read and write access to a personal data store. Hosting a dokieli document on a Solid server allows the author to edit the document directly from a space they control, and grant permissions to collaborators to do so as well. There are a number of open source Solid server implementations already, and users may host their own or choose a provider. There also exist Solid tools and libraries to help developers to build their own Solid-compliant data store.

Rich editing. The current implementation of dokieli makes use of the open source *Medium Editor* for the features one would expect from a WYSIWYG editor (figure dokieli-edit-menu). We have extended this with buttons for RDFa embedding, annotations and other interactions. The edit menu itself is loosely coupled with the rest of the application logic, so the choice of edit menu can be changed in future if better libraries become available.

Semantics and linking. The data format of dokieli documents is designed to encourage machine-readable description of ideas and knowledge, which becomes particularly powerful when these individual ideas are linked to each other with semantically meaningful relationships. A document author can assign a URI to concepts at any level, permitting reference to the document as a whole, a single phrase, or anything in between. Concepts within a single document can be related to each other with the appropriate RDF properties, as well as creating specific relations with external resources. dokieli takes a bottom-up approach to semantic content authoring; figure dokieli-rdfa shows our user interface for this, whereby the author selects some text, chooses the subject, predicate and object of a statement, and the application inserts the data as RDFa, and, if necessary, generates a new URI fragment to identify the subject or object.

In addition, metadata that does not easily fit into prose (for example, *Nanopublications* of data) can be embedded as a single block of Turtle, JSON-LD, or TriG (figure dokieli-nanopublication).

dokieli does not mandate the use of any particular RDF vocabularies, as the subject and content of an article determine how it is best described. By default, dokieli documents make use of *schema.org* for common terms for Web documents, the *SPAR Ontologies* for publishing and referencing, *Web Annotations* for comments and peer-reviews, *ActivityStreams* for social inbox notifications, the *Linked Data Platform* and *Solid* vocabularies for personal storage and user preferences, and the *ACL* vocabulary.

Authentication and access control. People can identify themselves to a dokieli document through WebID-TLS. Users with a browser certificate installed and a matching FOAF profile containing the certificate public key are authenticated against their own Solid server if found, with a fallback to a known Solid authentication endpoint. This allows the dokieli instance to write to any Solid-compliant data space which the user is authorised to write to. This is how users can, for example, save changes to a document, or store interactions such as comments.

Storage. Articles may be stored on the user's local filesystem or hosted on ordinary Web servers which can serve static HTML files, for example: university user pages, code repositories, personal or company web spaces, or any file hosting service.

4.2 Peer Review & Interactions

We support the rights of authors to own their data, and store and publish it where they feel most comfortable. This includes the social interactions and feedback users make around existing publications. Rather than centralising these interactions around the subject document, we took the decision to default to decentralisation of all content by allowing users to authenticate with their personal data space, and choose the location for their interactions at the point of making them. This gives rise to the need for a mechanism to notify the original author that their document has received some interaction. We achieve this by allowing document authors to specify an *inbox* for either their article as a whole, or any subsection with its own URI using the Solid `inbox` predicate. Inboxes are containers in a data space which may be appended to by anyone, and do not need to be on the same server as the document itself; if one article has multiple inboxes they can be distributed across as many data spaces as is convenient for the author(s). When an annotation is made, dokieli follows the appropriate `inbox` link and writes a notification there (see figure interactions-create for this process and listing 1 for notification contents). When the document is loaded, links are followed to all inboxes in order to retrieve interactions there have been notifications about, so that these can be displayed along with the document (figure interactions-display and figure dokieli-interactions).

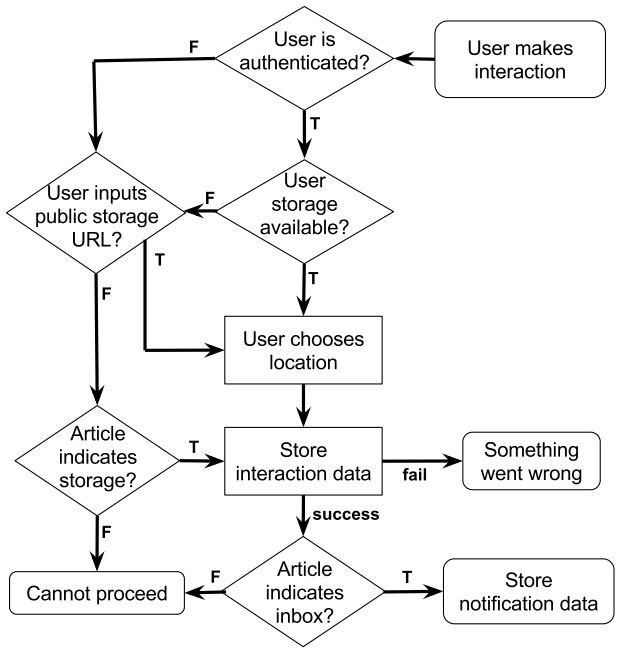

Figure 3: What happens when a user creates an interaction on an article.

```
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix schema: <https://schema.org/> .
@prefix solid: <http://www.w3.org/ns/solid/terms#> .
@prefix as: <http://www.w3.org/ns/activitystreams#> .
@prefix oa: <http://www.w3.org/ns/oa#> .
@prefix c: <http://creativecommons.org/licenses/by-sa/4.0/> .
<> a solid:Notification , as:Announce ;
  as:object <http://example.net/foo/abc123> ;
  as:context oa:hasTarget ;
  as:target <http://example.org/article#conclusions> ;
  as:updated "2016-01-24T00:00:00Z"^^xsd:dateTime ;
  as:actor <http://csarven.ca/#i> ;
  schema:license c: .
```

Listing 1: A Solid inbox notification.

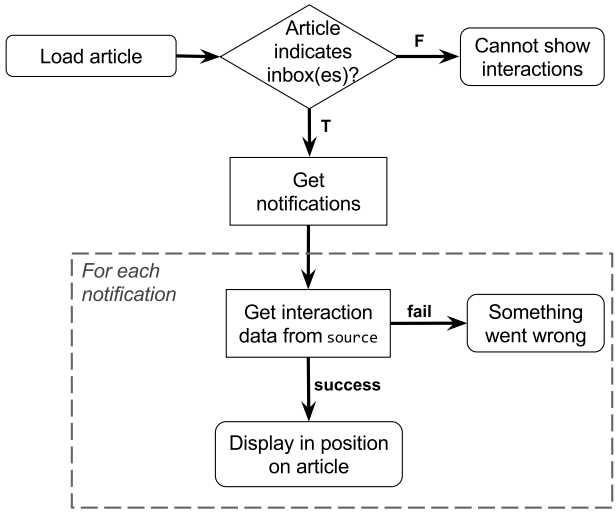

Figure 4: How interactions are displayed on an article.

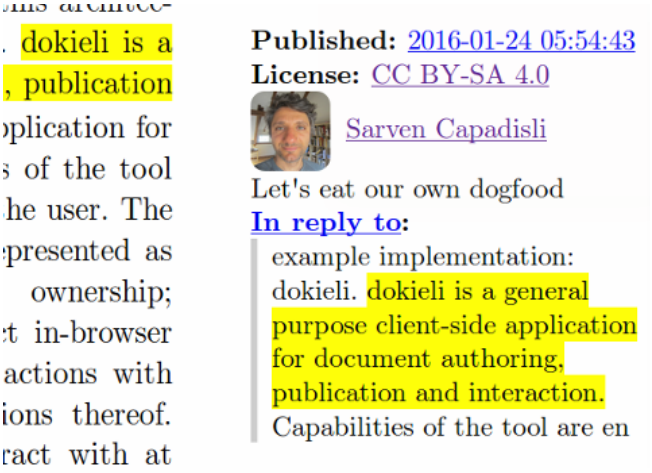

Figure 5: Interactions on a document.

Although providing enriched (meta)data is voluntary, provenance level data like the date on which the request was submitted, who by, and its license, can be purposed towards the verification process as well as for displaying. An example Solid inbox **Notification** is shown in Listing [1], in which **object** is the URL of an annotation, and **context** is the type of relation to the **target** (the conclusions section of an article).

dokieli itself does not offer a mechanism for authors to manage their notifications or choose which interactions appear, as other applications which are specialised for these tasks are under development as part of the Solid application ecosystem.

An author can also opt to allow anonymous interactions with their documents by pointing to a publicly writeable storage location in their own space, and store interactions on behalf of their audience.

A dokieli research article can contain a relationship e.g., `sioc:reply_of`, to the call for contributions. That is, every scholarly article can be written in context of the conferences' or journals' call for contributions e.g, the current article on Linked Research is a reply to *ACM Hypertext 2016 Calls for Contributions*, and contains an RDF statement as follows:

```
1 @prefix sioc: <http://rdfs.org/sioc/ns#> .
2 <http://csarven.ca/linked-research-scholarly-communication>
3   sioc:reply_of <http://ht.acm.org/ht2016/calls> .
```

Listing 2: Current article in reply to ACM Hypertext 2016 Call for Contributions (shown using the RDF Turtle syntax).

4.3 SemStats Call For Contributions

SemStats is a workshop that explores and strengthens the relationship between the Semantic Web and statistical communities. We have published all material on the site for the years 2013, 2014, and 2015 in HTML+RDFa using dokieli's tooling. It contains interlinks like:

- The *SemStats 2013* workshop is a `sioc:reply_of` to ISWC 2013 Call for Workshops
- The SemStats 2013 workshop also `schema:hasPart` *SemStats 2013 Call for Papers*

• SemStats 2013 Call for Papers then contains statements like: description, topic of interest, motivation, keywords, event data, related links, organising committee and affiliations.

4.4 SemStats CEUR Proceedings

In response to the SemStats 2013–2015 calls for papers, several papers have been submitted to the workshop. Those that were accepted for publication became part of the proceedings. Still, if one such paper had previously been published at its own URI, the proceedings version points there. We have published the SemStats proceedings at CEUR-WS.org (see, e.g., http://ceur-ws.org/Vol-1549/) and at the same time introduced a modernised version of the CEUR-WS.org proceedings template based on dokieli's HTML and CSS (but without JavaScript, which is not permitted by the CEUR-WS.org policy). Figure 6 demonstrates the interlinks between the peer-reviewed research article; *Linked Statistical Data Analysis*, SemStats 2013 call for contributions, ISWC 2013 call for workshops, and the proceedings of the SemStats workshop at CEUR-WS.org.

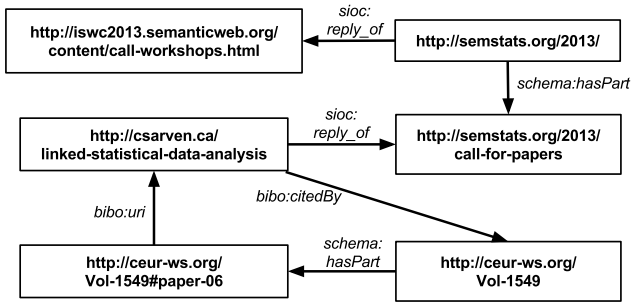

Figure 6: RDF statements interlinking a research article, call for contributions and workshops, and proceedings.

4.5 Ongoing And Future Work

There remains a number of features to add to our implementation to meet the requirements of the acid test.

Archiving and preservation. Organisations (or indeed individuals) who wish to make their resources available for archiving work can advertise this fact through linking to an inbox where *requests to archive* can be sent, optionally along with criteria (human- and machine-readable) which submitted material must meet. Requesting inclusion in an archive involves the same underlying mechanism as any kind of notification to an inbox (replying to a call, creating an annotation), though there are user interface optimisations we can make to streamline the process. Similarly for the managers of an archival collections, the technical mechanism for duplicating a document and creating a new URI exist in current 'Save As' functionality, however we can automate this for submissions to archives, whilst generating new semantic relations between the original documents and archival copies such as `sameAs` or `derivedFrom`.

Collections. Personal collections of work are similar to archives, proceedings and journals, except that one user is adding third-party documents to a collection over which they have control, so there is no request/accept process for something to be included. This is therefore similar to *bookmarking*, and it is likely that a *reference* to the original rather than a direct copy will be stored in the collection owner's space. dokieli functionality for managing collections of all kinds will also facilitate these kinds of personal collections, including the option to send a notification to the original author to let them know their work has been bookmarked. A particular use for this is for when a collection owner is writing their own academic article, and is able to auto-populate references and citations from their collections.

Profiles. An extension to dokieli which allows authors to pre-fill their affiliations and contact details permits automatic population of this data in new articles. Additionally, dokieli can automatically update an academic profile at the author's own domain whenever new publications are authored or feedback left on the work of others.

Notifications and feedback. We do not currently have good tooling for management of notifications and feedback on articles, particularly if feedback is able to be of different types (e.g. suggestions, questions, disagreements, notes) that maybe subsequently resolved by article authors. This may be in the form of an extension or separate, generalisable, application outside of the dokieli core.

Versioning. Much research is evolving, ongoing work, and we require a solution for better management and identification of snapshots of articles; in particular to associate feedback with a specific version of work which may be subsequently addressed, or to show which iteration of an article was accepted to a conference or journal.

5. RELATED IMPLEMENTATIONS

Taking the Linked Research requirements as a baseline and the acid test as a tool, we reviewed existing related systems with regard to how well they satisfy the requirements. Table 1 shows the results of profiling the systems against the requirements using the acid test.

Table 1: Comparison of publishing and consuming systems

System/tooling	DAaP	ACaA	PaA	PaP	SaSI	CaF	HaMR	IoS	EI	DI	AaPW	FaI	DVaM	AtA	IC	M	IMaRS
dokieli + Solid	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	
myExperiment		✓	✓		✓	✓	✓	✓	✓	✓	✓	✓				✓	
eLife		✓	✓		✓	✓				✓	✓	✓	✓		✓	✓	
OSF*		✓	✓		✓	✓				✓	✓	✓			✓	✓	
Google Drive		✓	✓		✓	✓				✓	✓	✓				✓	

System/tooling	DAaP	ACaA	PaA	PaP	SaSI	CaF	HaMR	IoS	EI	DI	AaPW	FaI	DVaM	AtA	IC	M	IMaRS
Authorea		✓	✓		✓	✓					✓	✓				✓	
Thinklab		✓	✓		✓	✓					✓	✓					

* Open Science Framework (OSF) is "one centralised location" which integrates with Dropbox, GitHub, Amazon Web Services, Box, Google Drive, Figshare, the Dataverse project, Mendeley

ACaA: Access control and attribution
AaPW: Authoring and publication workflow
AtA: Adaptation to audiences
CaF: Commentary and feedback
DAaP: Decentralised authoring and publishing
DI: Data integration
DVaM: Different views and media
EI: Entity identifiers
FaI: Feedback and interactions
HaMR: Human and machine-readability
IC: Interactive content
IMaRS: Impact metrics and reward system
IoS: Integration of semantics
M: Multimedia
PaA: Provenance and accountability
PaP: Persistence and preservation
SaSI: Sharing and social interactions

6. CONCLUSIONS

In this article, we have described requirements for exploiting the possibilities of digitisation and the Web for scholarly communication. The vision of Linked Research comprises concepts which can be realised by meeting these requirements, and we propose technical solutions for doing so. We demonstrate the feasibility of these proposals with a prototype implementation that covers several of the Linked Research concepts, and compare this coverage with existing tooling. We used this prototype, dokieli, for drafting and publishing this article.

Linked Research meets the criteria set forth in *The Five Stars of Online Journal Articles* [23] and we see this work as the first step on a larger research and development agenda. We envision an ecosystem of Linked Research compatible implementations and interfaces to existing systems (conference and journal management, editorial systems, OA repositories etc.). Ultimately, we hope that such developments will bring about a revolution in how scholarly knowledge is generated, published, shared and consumed.

7. ACKNOWLEDGEMENTS

The motivation and work on Linked Research is inspired by Marshall McLuhan, Ted Nelson, James Burke, and Tim Berners-Lee.

Special thanks to many colleagues whom helped one way or another during the course of this work (not implying any endorsement); in no particular order: Richard Cyganiak, Kingsley Idehen, Jodi Schneider, Paul Groth, Stian Soiland-Reyes, Raphaël Troncy, Anchalee Panigabutra-Roberts, as well as colleagues at MIT/W3C.

This research was supported in part by Qatar Computing Research Institute, HBKU through the Crosscloud project.

REFERENCES

[1] Borgman, C.: Scholarship in the Digital Age: Information, Infrastructure, and the Internet, 2007, ISBN 9780262026192, https://mitpress.mit.edu/books/scholarship-digital-age

[2] Berners-Lee, T.: W3C 1998, Webizing existing systems, https://www.w3.org/DesignIssues/Webize.html

[3] Procter, R., Williams, R., Steward, J., Poschen, M., Snee, H., Voss, A., Asgari-Targhi, M.: Adoption and use of Web 2.0 in scholarly communications. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 2010(368):4039-4056, DOI: 10.1098/rsta.2010.0155, http://rsta.royalsocietypublishing.org/content/368/1926/4039

[4] Licia, C., Cassella, M.: Scholarship 2.0: analyzing scholars' use of Web 2.0 tools in research and teaching activity. Liber Quarterly 23.2 (2013): 110-133, https://www.liberquarterly.eu/articles/10.18352/lq.8108/

[5] Neylon, C., Wu, S: Open Science: tools, approaches, and implications. Pac Symp Biocomput. 2009:540-4, http://psb.stanford.edu/psb-online/proceedings/psb09/workshop-opensci.pdf

[6] Piwowar, H. A., Day, R. S., Fridsma, D. B.: Sharing Detailed Research Data Is Associated with Increased Citation Rate, PLOS ONE, 2007, DOI: 10.1371/journal.pone.0000308, http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0000308

[7] Ensor, P.: The Functional Silo Syndrome, AME Target (1988):16, http://www.ame.org/sites/default/files/documents/88q1a3.pdf

[8] Work, S., Haustein, S., Bowman, T. D., Larivière, V.: Social Media in Scholarly Communication, Canada Research Chair on the Transformations of Scholarly Communication, 2015, http://crctcs.openum.ca/files/sites/60/2015/12/SSHRC_SocialMediainScholarlyCommunication.pdf

[9] Kuhn, T. S.: The Structure of Scientific Revolutions, University of Chicago Press, 1962, ISBN 9780226458113

[10] Wagner, W., Steinzor, R.: Rescuing Science from Politics: Regulation and the Distortion of Scientific Research, Cambridge University Press, 2006 p. 224, ISBN 9780521855204

[11] Wicherts, J. M.: Peer Review Quality and Transparency of the Peer-Review Process in Open Access and Subscription Journals, 2016, DOI: 10.1371/journal.pone.0147913, http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0147913

[12] Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.): FORCE11, https://www.force11.org/datacitation

[13] Capadisli, S., Auer, S., Riedl, R.: This ‘Paper’ is a Demo, ESWC Satellite Events (2015), http://csarven.ca/this-paper-is-a-demo

[14] Ahalt, S., Carsey, T., Couch, A., Hooper, R., Ibanez, L., Idaszak, R., Jones, M. B., Lin, J., Robinson, E.: NSF Workshop on Supporting Scientific Discovery Through Norms and Practices for Software and Data Citation and Attribution. Technical Report. National Science Foundation, USA, 2015, http://dl.acm.org/citation.cfm?id=2795624

[15] Capadisli, S., Riedl, R., Auer, S.: Enabling Accessible Knowledge, CeDEM (2015), http://csarven.ca/enabling-accessible-knowledge

[16] Library Linked Data in the Cloud: OCLC's Experiments with New Models of Resource Description. Carol Jean Godby, Shenghui Wang, and Jeffrey K. Mixter Synthesis Lectures on the Semantic Web: Theory and Technology 2015 5:2, 1-154, http://www.worldcat.org/oclc/909811018

[17] Osborne, F., Scavo, G., Motta, E.: Identifying diachronic topic-based research communities by clustering shared research trajectories, Research Track, ESWC (2014), http://oro.open.ac.uk/39666/3/ESWC2014_CR

[18] Berners-Lee, T.: Cool URIs don't change, W3C (1998), https://www.w3.org/Provider/Style/URI.html

[19] Rosenthal, D. S. H.: What Could Possibly Go Wrong? BIBLIOTHEK – Forschung und Praxis 2015; 39(2): 180–188, DOI 10.1515/bfp-2015-0022, http://www.lockss.org/locksswp/wp-content/uploads/2015/06/bfp-2015-0022-1.pdf

[20] Markman, C., Zavras, C.: BitTorrent and Libraries: Cooperative Data Publishing, Management and Discovery, Volume 20, Number 3/4, 2014, doi:10.1045/march2014-markman, http://www.dlib.org/dlib/march14/markman/03markman.html

[21] Kuhn, T., Chichester, C., Krauthammer, M., Dumontier, M.: Publishing without Publishers: a Decentralized Approach to Dissemination, Retrieval, and Archiving of Data, ISWC 2015, http://iswc2015.semanticweb.org/sites/iswc2015.semanticweb.org/files/93660593.pdf

[22] Capadisli, S., Guy, A., Auer S., Berners-Lee, T.: dokieli: decentralised authoring, annotations and social notifications, 2016, http://csarven.ca/dokieli

[23] Shotton, D.: The Five Stars of Online Journal Articles, D-Lib Magazine (2012), http://www.dlib.org/dlib/january12/shotton/01shotton.html