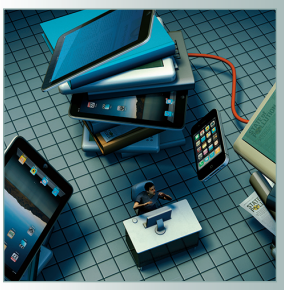

Addressing Information Overload in the Scientific Community

The authors present Liquid Journal, a dissemination model and website that extends from the notion of a traditional scientific journal to overcome the problem of information overload in the scientific community. They detail Liquid Journal's concepts, methods, and supporting platform. They also focus on the issues related to having access to a plethora of relevant scientific content, such as narrowing down the discovery process to reliable sources.

Marcos Baez, Aliaksandr Birukou, Fabio Casati, and Maurizio Marchese
University of Trento

The Web has opened a whole world of possibilities for how we create, evaluate, and disseminate scientific knowledge. We can now publish preprints in online archives (such as arXiv) or simply post our papers on webpages. Furthermore, "papers" are not the only unit of scientific dissemination. Data, comments, scientific experiments, and even blogs can now be shared and considered a form of scientific contribution that can help other scientists in their work. This means that today we have a large scientific community easily presenting a vast, and rapidly evolving, set of scientific contributions. An implication of this is that, while in the past the scarce resource for scientific dissemination was printing, now it's *attention*. The obstacle to dissemination is how to find interesting and relevant information (for readers) and to make the work visible in the sea of virtually infinite information (for authors).

An additional and somewhat puzzling problem of information overload is that, with so much information available, we'd like to be able to broaden our horizons, but it's difficult. For example, we'd like to be able to search for contributions on the effectiveness of peer review in many different domains (as this problem is indeed studied in different areas). However, having this much information results in having to narrow down what we read as opposed to broadening it. We experience this in everyday life: having a digital video recorder (such as a TiVo) makes a wide set of TV programs available to us, but then we tend to watch/record what we know we like, and are less encouraged to look for new programs. We can observe the same effect in science,[1] given that we tend to keep looking into the sources we're familiar

1089-7801/10/$26.00 © 2010 IEEE

with, thereby missing a plethora of potentially interesting and relevant contributions.

Today, only a few tools are at our disposal to leverage this richness of information while handling the overload (we review these tools in the sidebar on p. 34, "Related Work in Information Dissemination and Overload"). When we search for contributions, we still tend to look for papers or articles, and one option is to look at a collection of papers or articles indexed by services such as the Digital Bibliography & Library Project (DBLP) or CiteSeer. This is somewhat useful, but it doesn't solve the problem: we're still limited to what's indexed, to papers or articles (and only those that have been published), and to a narrow selection (such as services only available in computer science for the most part). Despite this narrowness, we're still likely to be overloaded with the results. An alternative approach is to use a generic Google search, which isn't tailored to finding scientific contributions, or Google Scholar, Google's specific search engine for scientific contributions, but the results aren't often as helpful as when we search the Web for other purposes. Furthermore, even when we find something we like, we can navigate to related content only via citations, inserted by the authors at the time of writing.

With this in mind, we propose the notion of *liquid journals* (an evolving collection of interesting and relevant links to scientific contributions available on the Web) as a way to overcome information overload in scientific publications. Their underlying principles are

- leveraging the same (large) community of scientists that creates the overload problem with the opportunity to collaborate in filtering and prioritizing the information;
- enabling a dissemination and consumption model that naturally reduces the noise right at the source;
- having a set of metrics that mitigate the overload and encourage "good behaviors" for science, such as early sharing and providing feedback; and
- facilitating readers in linking knowledge, which will support other users' subsequent searches and navigation through related content.

Liquid journals put these principles to work through concepts, methods, and ultimately tools. In this article, we detail our usage model along with its derived metrics and a sample website — called Liquid Journal (see http://liquidjournal.org).

Basic Concepts of the Liquid Journal

Liquid Journal builds on the idea of a model for scientific contributions that's designed to facilitate the search for — and navigation of — scientific information of interest. We see scientific contributions as structured, evolving, and multifaceted objects. Specifically, we see scientific content as something that we want to search within and help assess and disseminate by spatially representing it as scientific resources organized as a set of nodes in a graph that authors, editors, or even readers can connect or annotate. The reason for connections, and hence for modeling resources as a graph, is to capture several kinds of dependencies or relationships among them (or between resources and people or other entities).

To illustrate these concepts, Figure 1a shows our research group's work on evaluation metrics and peer review. We started this line of research within the context of a project deliverable called D3.1. This deliverable contained a review of the state of the art, experiments, analysis, and presentation of the results. We delivered the results in two releases — D3.1v1 and D3.1v2 (see https://dev.liquidpub.org/svn/liquidpub/final/Year1/LP_D3.1.pdf and https://dev.liquidpub.org/svn/liquidpub/wp3/d3.1/v2/) — and we plan to produce in the near future a third version. These releases are captured by relations that let us specify when a particular scientific resource is completely new, or if it's the latest iteration of a previous one.

Recently, we achieved some interesting results that we wanted to communicate, so we took some work from the second version of our deliverable and produced a technical report called "Is Peer Review Any Good?" (see http://eprints.biblio.unitn.it/archive/00001654/). This type of spin-off is captured by different branches in the graph, to show the timeline of research. This kind of graphical representation is helpful when we want to expand our search of a particular scientific resource (such as "Is peer review any good?"). We can see, then, that this resource has many representations (see Figure 1b). These alternative representations are different views of the same resource, such as slide sets, a technical report, and a conference paper.

We can also see how this scientific resource is semantically related to other entities. Figure 1c illustrates the use of particular data and experiments (for example, conference review data and the code that processes them). These links help readers, editors, and authors connect and describe relationships among resources.

We define these relationships because it helps leverage the power of the community to build scientific dissemination knowledge — that is, knowledge that can help annotate and relate resources above and beyond what authors would do. In other words, people generate knowledge that helps in organizing and finding scientific resources. This is sometimes called *second-order knowledge*, which we believe is as important in supporting scientists' work ("standing on the shoulders of giants") as first-order knowledge, provided by authors and publishers.

Formally, we define the space of scientific resources as $\Sigma = <SR, E, L, A>$, where

- SR is a set of resources where $r = <id, uri, ct, cf>$ are the individual scientific resources. Here, id denotes the universal identifier for the resource; uri points to the resource as available on the Web; ct is the resource's content type and can take values such as paper, video, slide set, dataset, and experiment; and cf is the content format (which can be .pdf, .pptx, and so on). Because we consider journals as a way to create or at least disseminate knowledge, they're also resources.
- E is the set of entities that create, access, relate, annotate, or certify resources. These can be people or institutions (including certification agencies).
- L denotes a set of links $l = <e_s, e_t, lt, u, un>$ representing relations among resources or between resources and entities (from source e_s to target e_t). Besides the objects they relate, they're essentially characterized by a type lt (such as "next version of"), by the users $u \in E$ that created it, and by the users or agencies that *endorse* it, if any, $un \in E$.
- A denotes a set of annotations $a = <e, at, v>$ that can be attached to resources or an entity e. Annotations can be of a certain type at (such as tags, flags, and comments), and carry a value v (such as "good example of state of the art").

While Liquid Journal lets anybody create any kind of relation, it assumes and leverages specific relation types, to which it assigns an agreed semantic (and also graphical interaction patterns in the Liquid Journal interface):

- *Structural relations* represent arbitrary relationships between contributions, where the relationship is described by annotations. For example, a paper can *report on* a dataset in that it describes results of experiments on that dataset. We depict examples of such relations in Figure 1c.
- *Temporal relations* (such as the *next_version_of* relation) model the evolution of a resource, be it a paper, dataset, or anything else. This is a natural behavior of research dissemination, where for example we write a preliminary version of a paper and then extend or refine it. Or, we clean or add more data to a dataset. Figure 1a also shows that evolution can follow a line (as in multiple versions of our project's D3.1) or branch (from one deliverable, we then derive a paper or technical report).

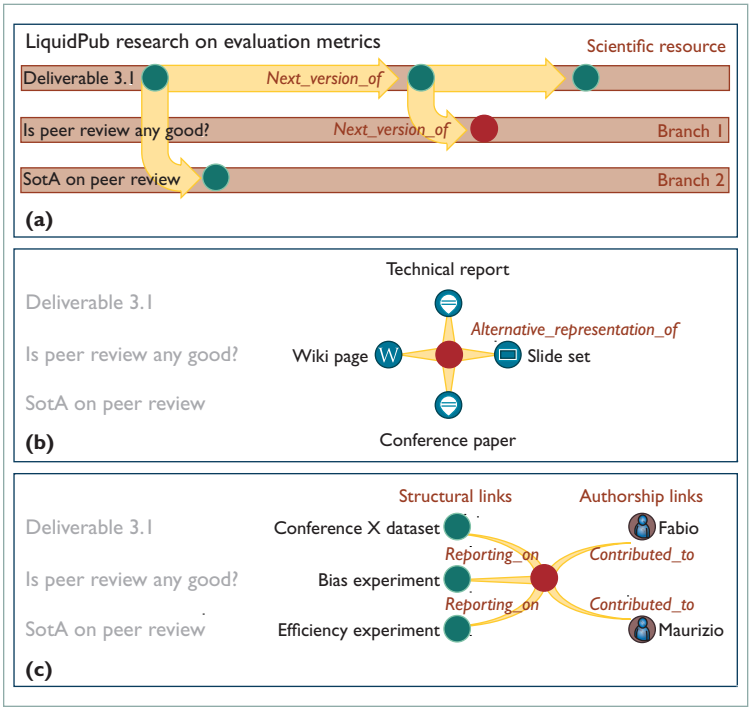

Figure 1. An example of graphs of scientific resources on evaluation metrics and peer review. These graphs capture the dependencies and relationships between resources and people or other entities. We can see example graphs of (a) scientific resources connected by the next-version-of *relation, (b) different representations for the same resource, and (c) other general relations, such as authorship or structural relations.*

Related Work in Information Dissemination and Overload

Despite the progress in dissemination models, the current model of publishing and evaluating scientific contributions remains almost the same. Novel models such as the deconstructed model[1] and the overlay journal[2] introduce interesting ideas that should be further elaborated to be applied in and get benefits from Web 2.0. These models are still constrained to the traditional notion of paper, thus other contribution types remain hidden. The social aspect for these approaches, however — the study of behaviors that are good for science, such as early feedback, sharing, and collaboration — remain unexplored. More importantly, none of the models tackle or offer mechanisms to face the problem of attention. All these issues also affect the evaluation, which continues to be based mostly on citation-based papers or articles,[3,4] thereby leaving out other aspects of research productivity.

The social Web has made new forms of collaboration possible. Prominent examples are social bookmarking services that let users share interest within communities. CiteULike, Mendeley, Zotero, and Connotea are examples of social bookmarking services that focus on sharing and organizing academic references. These tools come with social tagging features that let people collaboratively tag content. Thus, these tools provide storing, sharing, and tagging of references to publications via shared collections and groups.

Tools for sharing and collaboration offer a promising direction. These systems provide a foundation of results for further study in the scientific domain regarding collaboration. However, these systems are only a short-term way to collaborate until a formal and complete knowledge-dissemination model is established. Moreover, taking technical aspects apart, one disadvantage of these services is that they rely on active users — that is, users who inject content into the system. Thus, the discovery is limited to what's already there. Our model builds on some social features of these systems, but provides a complete model of dissemination designed especially to overcome the scientific domain's dissemination overload.

Search is a common service on the Web, so search engine technology has been explored and applied to scientific content.[5] Specialized search engines such as Google Scholar and CiteSeer have been developed for searching for papers, articles, and books across multiple repositories using crawling techniques and protocols. Using another approach, the Bielefeld Academic Search Engine (BASE; www.base-search.net) indexes the metadata from repositories that implement the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). In addition to what the user can provide as input to the search (such as keywords), implicit preferences and collaborative filtering have also been used for yielding content that users might like.[6] This has led to proposals of relevance and diversity algorithms that try to balance user preferences and diversity.[7] In the academic domain, many studies have explored peers' recommendations of papers or articles.[8]

Thus, academic search engines provide only a partial view of the scientific contributions dispersed over several sources on the Web. They don't capture user preferences and lack of proactive behavior. Users need to know what and how to search to get the content they want, and when they do find an interesting resource, the navigation and exploration is limited. General approaches provide a foundation, but their use in the scientific domain needs to be modeled for the broader notion of scientific contribution, as well as other special issues in the scientific domain (such as ranking). In our approach, we rely on a model that provides semantic relations that we can exploit to explore and discover new, similar, and related scientific resources.

References

1. J. Smith, "The Deconstructed Journal — A New Model for Academic Publishing," *Learned Publishing*, vol. 12, no. 2, 1999, pp. 79–91.
2. S. Pinfield, "Journals and Repositories: An Evolving Relationship?" *Learned Publishing*, vol. 22, no. 3, 2009, pp. 165–75.
3. J. Hirsch, "An Index to Quantify an Individual's Scientific Research Output," *Proc. Nat'l Academy of Sciences*, vol. 102, no. 46, 2005, pp. 16569–16572.
4. M. Krapivin, M. Marchese, and F. Casati, "Exploring and Understanding Citation-Based Scientific Metrics," *Advances in Complex Systems*, vol. 13, no. 1, 2010, pp. 59–81.
5. N. Lossau, "Search Engine Technology and Digital Libraries," *D-Lib Magazine*, vol. 10, no. 6, 2004; www.dlib.org/dlib/june04/lossau/06lossau.html.
6. J. Schafer et al., "Collaborative Filtering Recommender Systems," *The Adaptive Web*, LNCS 4321, Springer, 2007, pp. 291–324; doi:10.1007/978-3-540-72079-9_9.
7. M. Hadjieleftheriou and V.J. Tsotras, eds., "Special Issue on Result Diversity," *Bulletin of the Tech. Comm. on Data Eng.*, vol. 31, no. 4, 2009; http://sites.computer.org/debull/A09dec/issue1.htm.
8. D. Parra and P. Brusilovsky, "Evaluation of Collaborative Filtering Algorithms for Recommending Articles on CiteULike," *ACM Conf. Hypertext and Hypermedia, Workshop Web 3.0: Merging Semantic Web and Social Web*, 2009; www.slideshare.net/denisparra/evaluation-of-collaborative-filtering-algorithms-for-recommending-articles-on-citeulike.

- *Representation relations* let us model the multifaceted aspect. For example, a paper can have associated slides and datasets, and therefore be deemed as a complex, multifaceted artifact, including artifacts that encode (part of) the same knowledge but have a different representation. Figure 1b illustrates an example of this type of relation.
- *Authorship relations* denote who contributed to the creation of the resource. An annotation of this relation would qualify the contribution (such as "design of the experiment").

- *Dissemination relations* denote usage by means of the Liquid Journal model. For example, they include the appearance of a resource in a journal, subscription to a journal, and sharing of a resource.

This model for resources reduces overload because it clusters contributions into research lines (which are themselves resources) and then lets users navigate through contributions in those lines, as well as evolutions, presentations, and other related resources. Although it's outside this article's scope, this approach makes it easier to fairly attribute credit to contributions or authors by making explicit a contribution's incremental nature and to assign an indirect reputation to resources because they're linked by another (reputed) resource — much like what Google's PageRank does for webpages.

As we previously noted, we consider a liquid journal an evolving collection of interesting and relevant links to scientific contributions available (whether freely or not) on the Web. Considering these journals as *collections of links* means that the journals do not own the contributions. We assume that the contributions are posted elsewhere as webpages, traditional journals, and so on — and thus they're independent of their appearance in a journal. (However, scientific contributions can point to reliable archives and, in general, sources that ensure long-term persistence.) Many journals can refer to the very same contribution. This "appearance" of contributions in journals is important for measuring and determining their *interestingness*.

The links in a journal (which define its content) can be decided by the editor, who picks them one by one, or defined by a Web search through the liquid journal's engine, where the results depend on the resources' interestingness. The editor can then refine the search result and "snapshot" it (resulting in an *issue* of a liquid journal), or the journal can adopt a continuous model in which it's essentially a Web search, and the result evolves naturally and continuously as new content becomes available or the values of metrics for existing contributions make them qualified for the defined journal.

The rationale behind this model is that we see journals as a mechanism for people to find and share interesting and diverse content, for themselves or for their research group. This was also the original motivation at the birth of the scientific journal's paper model around the 17th century. It's also why we believe that our new model correctly maintains the name *journal*. While doing this — while running a liquid-journal-enabled search for Web content — and while refining the results and sharing the most interesting contributions with our colleagues, we do a service to our team; but we're also acting as filters in that we implicitly rate contributions. Hence, we're also doing a service to the community. Liquid journals essentially put the community itself to work as content selectors while having people perform activities they need to do anyhow, such as look for content and share interesting findings with their team. It's like capturing the interestingness that people perceive from the result of a Web search and using this as a way to rate content and therefore separate more interesting contributions from the rest.[2]

Usage Model and Metrics

Liquid journals aim at providing tailored scientific content by bringing interesting scientific contributions. People fill their liquid journals in various ways: they can add content they stumble on by emailing a .pdf file to the liquid journal engine (analogous to "digging" an item), or even by taking a picture of a paper with their phone. They can also add a work in progress, such as a Google document (see the demo videos at www.youtube.com/user/liquidjournals for details). This is intended to mimic what we do today to keep track of interesting contributions. The actual content, however, isn't in the journal. A liquid journal is a collection of links, and as such, it relies on the actual sources and on the editor's ability to access those sources. Thus, access permissions are always based on the reader's permissions and on what the link's source allows.

A value proposition of a liquid journal is that editors and readers will provide knowledge that can help connect and assess scientific contributions. This happens in three ways, all supported by the Liquid Journal interface:

- Editors implicitly evaluate resources by publishing them in their journal(s).
- Readers implicitly evaluate resources by sharing them with their team. For example, a professor or a doctoral student can share

papers or articles that they think are interesting within their team.
- Readers, editors, and authors provide knowledge by linking and annotating resources. For example, a reader can state that paper *P*1 reports results of experiment *E* over dataset *D*, and extends the inital results of *P*2. They can also state that paper *P*3 offers a nice literature review.

The third action provides information that's useful for navigating from a resource to related resources, and therefore for finding related information, as shown in Figure 1c.

With the first two actions, the scientific community collectively establishes what's worth reading. Feedback in this form isn't intrusive, and it can be useful for editors or readers. This work of selecting and sharing knowledge is what we do every day. What the Liquid Journal tries to add is to capture this information by making it easy and convenient for each of us to select and share resources and then use the collective (implicit) opinions people have expressed to select and share content. In other words, by giving scientists a tool to collect, organize, and share interesting scientific resources, we have a way to assess the interestingness of such resources, and consequently a way to filter interesting knowledge and help manage the information overload. Furthermore, expanding the reach of metrics to other types of content and activities will let us look into other aspects of researchers' productivity. For example, we can explore how to reward people for sharing good ideas (such as by posting them in a blog), selecting and creating good collections of contributions, and also giving constructive feedback. Traditional metrics not only can't provide such insights, but they're still based on citations, which have been shown to have flaws.[3]

Liquid Journal's conceptual model also provides the information to capture these aspects in the dimensions of scientific contributions, subscription links, structural links that make contributions appear in journals, and usage information (such as tags, forwarding, and sharing). The traditional model doesn't cover all these rich information-gathering aspects.

From an evaluation perspective, we see the main contribution of this work in providing the basic information for evaluating all sorts of resources based on community opinions implicitly provided. Out of these, we can develop many new metrics, just like many citation-based metrics have popped up now that it's possible to compute citations automatically. A trivial approach involves counting the number of journals in which a resource appears, or the number of people who share it or tag it. Nadine Osman and her colleagues provide a more sophisticated example,[4] where opinions, tags, journal selections, and other actions that can be expressed via (and recorded by) a liquid journal contribute to a resource's reputation. This is the algorithm currently integrated with the Liquid Journal platform.

However, because it's infeasible to provide a unique (and accepted) magic formula that captures all of these aspects, we focus on providing the guidelines that will govern the instantiation of particular derived metrics. Indeed, we believe it's up to the community to decide what counts within it. We're developing this concept with the metric uCount (in joint collaboration with the Institute for Computer Sciences, Social Informatics, and Telecommunications Engineering; www.icst.org) that, as the name suggests, captures both the fact that everyone in the community counts and that everyone's involved in the process of defining what counts in his or her specific community. The idea is that anybody can then decide which metric formula to use to filter out potential resources of interest when searching for Web content.

Architecture

Designing and implementing an infrastructure for supporting the Liquid Journal model requires solutions and strategies for

- managing the journal's process;
- journal creation, evolution, consumption, and sharing;
- access to scientific content in the Web;
- computing the reputation of contributions (for ranking); and
- projecting these features onto a user interface.

The Liquid Journal architecture relies on specialized components designed for each of these aspects. We illustrate these components in Figure 2a.

Our current Liquid Journal site provides a

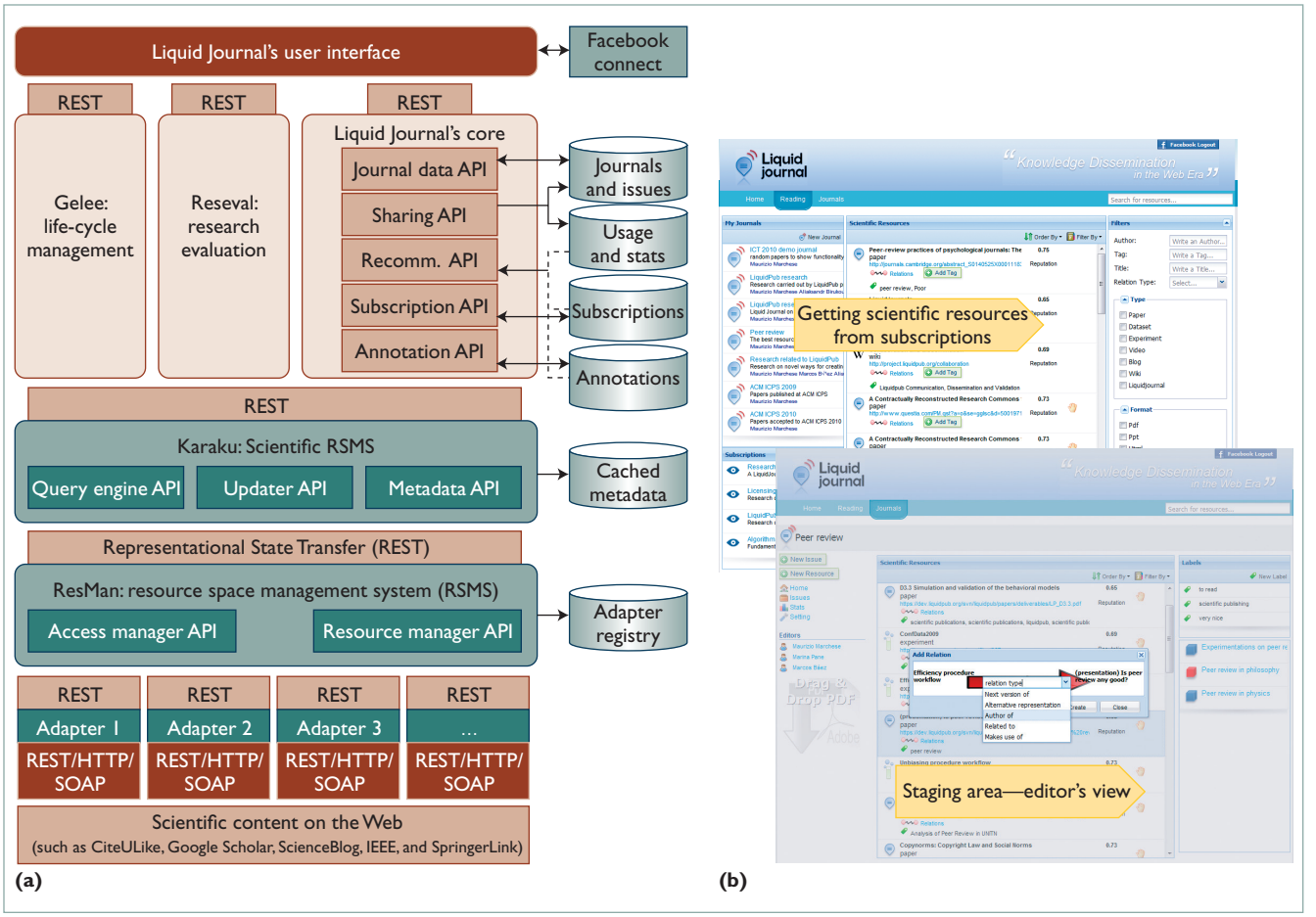

Figure 2. The Liquid Journal *architecture's components. We can see (a) the back-end architecture, and (b) the front-end screenshots.*

view of the scientific content available on the Web. Because many scientific contributions fall outside traditional sources (such as digital libraries) where standards can be applied, the infrastructure requires an access layer that provides the necessary abstractions for accessing and searching content on the Web. To address this requirement, we rely on the abstraction of a resource space-management system (RSMS)[5] applied to the scientific domain.

The ResMan system (see http://project.liquidpub.org/resman), a prototype implementation of an RSMS, provides a uniform access layer to resources available on the Web. It abstracts applications on top of the underlying Web services' heterogeneity. The approach the system follows is to rely on adapters — that is, components that map the specifics of different and incompatible services to a common and uniform protocol.[6]

On top of ResMan, the abstraction of a scientific RSMS named Karaku (http://project.liquidpub.org/karaku) provides a common and extensible conceptual model specific for scientific resources, and a set of basic services for searching and operating on these resources. A core module is the Updater, which functions as a crawler over scientific sources and extracts resource metadata. This lets us push resources into the system the same way that users interface with their phone or a Web interface. On these architectural foundations, the Liquid Journal's core component builds the services that support the model introduced in this article.

Liquid journals let users define their own process and, to this end, the architecture also includes a life-cycle management component, the Gelee system.[7] The back end is completed by a research evaluation tool (Reseval; http://project.liquidpub.org/reseval), an extensible tool for computing metrics for contributions and papers (and any other user-defined entity). In this context, the tool takes information about scientific

entities from Karaku and applies the algorithms for computing metrics. Thus, we can view liquid journals as domain-specific mashups that let users define the content, process, and metrics.

Services are important in our architecture, but to fully exploit them, we must provide an effective Web interface that facilitates journal definition, search, content consumption, and sharing. In our approach, we pay special attention to this issue, and we're developing a rich Web application on top of the core components (see Figure 2b). It's also possible to access the Liquid Journal application using Facebook's social network log in and password. We did this with the goal of facilitating sharing and making it easier for people to use and connect with the system.

We developed the Liquid Journal model in cooperation with Springer and other partners of the LiquidPub project. Currently, it's being deployed as part of ICST — and, as such, made available to a large community of users. We hope that Liquid Journal will provide support for scientists who would like to collaboratively collect, organize, and share relevant content. Another target audience is the reader who can create knowledge by tagging, commenting, and linking different contributions in liquid journals. Such user-created knowledge is a basis for novel metric models and for approaches that automatically map contributions, authors, and venues to communities. We intend to incorporate search functionalities based on communities and user-created knowledge in future iterations. Further details are available at http://project.liquidpub.org/research-areas/liquid-journal.

Acknowledgments

This work was supported by the European Union Information and Communications Technology (EU ICT) project LiquidPublication, under FET-Open grant number 213360.

References

1. J. Sandweiss, "The Future of Scientific Publishing," *Physical Rev. Letters,* vol. 102, no. 19, 2009; doi:10.1103/PhysRevLett.102.190001.
2. D. Kelly and J. Teevan, "Implicit Feedback for Inferring User Preference: A Bibliography," *ACM Special Interest Group on Information Retrieval Forum*, vol. 37, no. 2, 2003, pp. 18–28.
3. A. Haque and P. Ginsparg, "Positional Effects on Citation and Readership in ArXiv," *J. Am. Soc. Information Science and Technology*, vol. 60, no. 11, 2009; doi:10.1002/asi.21166.
4. N. Osman et al., *Credit Attribution for Liquid Publications,* LiquidPub deliverable, 10 June 2010; https://dev.liquidpub.org/svn/liquidpub/papers/deliverables/LP_D4.1.pdf.
5. M. Baez and F. Casati, "Resource Space Management Systems," *Proc. European Conf. Web Services,* IEEE CS Press, 2009, pp. 3–4.
6. B. Benatallah et al., "Developing Adapters for Web Services Integration," *Proc. Center for Advancement Informal Science Education*, Springer, 2005, pp. 415–429.
7. M. Baez, F. Casati, and M. Marchese, "Universal Resource Lifecycle Management," *Proc. Int'l Conf. Data Eng.*, IEEE CS Press, 2009, pp. 1741–1748.

Marcos Baez is a doctoral student at the University of Trento, Italy. He's actively involved in the LiquidPub project, where he works on models, tools, and algorithms for improving the way scientific research is disseminated. His research interests include data spaces, mashups, and collaborative aspects of social networks. Baez has a master's degree in informatics engineering from the National University of Asuncion, Paraguay. Contact him at baez@disi.unitn.it.

Aliaksandr Birukou is a postdoctoral student at the University of Trento. His current research interests include improving the way science works, patterns, recommendation systems, culture, and compliance management. Birukou has a PhD in information and communication technologies from the University of Trento. Contact him at birukou@disi.unitn.it.

Fabio Casati is a professor of computer science at the University of Trento. His research interests include exploring solutions for Web services and business process management. Casati has a PhD in information engineering from the Politecnico di Milano. Contact him at casati@disi.unitn.it.

Maurizio Marchese is an associate professor of computer science at the University of Trento and the principal investigator in the LiquidPub project. His research interests include the design and development of service architectures in distributed systems, as well as the analysis, development, and integration of services to support and enhance scientific knowledge creation and dissemination processes. Marchese has a PhD in physics from the University of Trento. Contact him at marchese@disi.unitn.it.