

The Embryo Project and the Emergence of a Digital Infrastructure for History and Philosophy of Science

Manfred D. Laubichler, Jane Maienschein, and Grant Yamashita

Abstract. The international collaborative Embryo Project¹, housed at Arizona State University is developing an online open access encyclopedia covering 2500 years of embryo research in multiple varied contexts. It is also one of the core projects that make up a new initiative to bring the tools of the bioinformatics revolution in the life sciences to the areas of history and philosophy of science (HPS)². This paper discusses the structure and accomplishments of the Embryo Project in the wider context of digital HPS with a special emphasis on ongoing work in developing a cyberinfrastructure for HPS.

Over the last two decades the life sciences have experienced unprecedented growth and have developed a new explanatory paradigm of systems biology (Kitano 2001; Alon 2006). These conceptual developments have been based on a shift away from simple causal explanations based on clearly identifiable factors, such as individual genes, and towards an emphasis on multiple factors and their interactions. Technological advances, especially in the area of bioinformatics, have played a central and crucial part in these conceptual changes (see for instance the National Center for Biotechnology Information for an overview of the tools and principles

¹ <http://embryo.asu.edu>

² <http://www.digitalhps.org>

of currently deployed in bioinformatics)³. As a consequence, the life sciences are now to a large degree information-based with the relevant information stored in both centralized and distributed databases. Sophisticated search algorithms and queries based on conceptual models and ontologies (in the computer science designation of the word), standardized annotation practices and new generations of relational databases connecting different kinds of data are the foundation of these new forms of life science research (Ouzounis and Valencia 2003).

Scholars in the Science Studies community (history, philosophy, and sociology of science) have always emphasized complex explanations of historical events, which are mostly presented as historical and richly contextual narratives and are thus always the end result of years of individual scholarship. What the science studies community has not yet embraced are the enormous benefits of the informatics revolution that has transformed the life sciences with respect to the organization of multiple forms of complex data, shared access to these data, searches in distributed relational databases that are organized around standardized practices of database management, and the possibilities of digital workbenches for collaborative and distributed research. All these developments have also contributed to robust cyberinfrastructure, which has changed the ways biologists go about their research (Ouzounis 2002, Lesk 2008).

In other words, the science studies community is missing out on new ways to conduct and organize research and to store, distribute and analyze data. One of the main consequences of the bioinformatics revolution has been the possibility of large-scale and comparative analysis of data and the integration of detailed experimental research with readily available points of comparison. This strategy has facilitated a bottom-up approach that allows biologists to find patterns of increasing generality. Insofar as one goal of the science studies community is to better understand both individual sciences as well as science at large in its various contexts (technological, theoretical, historical, social or political), it too will have to move beyond the particular and focus on general patterns wherever these do exist, a goal greatly facilitated by the tools of the informatics revolution.

Unfortunately, current practices of both training and research have left researchers in the history and philosophy of science, and the humanities as a whole, ill-prepared and ill-equipped to take advantage of the new developments that have transformed research in the life sciences over the last two decades (Cantara 2004). Within the science studies community the history and philosophy of biology provides an excellent entryway into examining how to incorporate informatics technologies into its research because it bridges studies in the life sciences (which have developed the informatics technologies) and the humanities.

To give just one concrete example of the scope of the challenges, John Tyler Bonner, Professor Emeritus at Princeton University and doyen of the slime mold research community, mentioned at a recent workshop on morphogenesis at

³ <http://www.ncbi.nlm.nih.gov/>

Arizona State University (ASU) that for most of his career he could read and keep track of every paper published on his preferred research subject, slime molds, but that today he is no longer able to come even close to doing that even though slime mold research is still a small field by anybody's standards (Bonner 2008). And slime mold research is still a small field by anybody's standards. The explosion of information produced by an increasingly globalized research community is simply staggering.

While all academics have experience in what may be appropriately called information triage, it is almost impossible to keep up without the help of newly developed information technologies that affect all areas of knowledge production, storage and dissemination (Renar and Palmer 2009). But before individual areas of research can take full advantage of these new possibilities we not only need to develop additional technologies and solve problems related to standards, we also need to change the culture in respective fields, such as the humanities and science studies, in order for them to embrace these new options (Martin 2007).

In this context the ASU-based collaborative Embryo Project (EP) has already provided unique insights into the possibilities of a digital approach to history and philosophy of science (HPS). The Project is also revealing numerous and varied technical and sociological challenges that still need to be addressed before serious digital HPS can come to fruition. Here we first introduce the Embryo Project and then give a brief overview of the possibilities and challenges of a digital framework for history and philosophy of science.

The Embryo Project

The Embryo Project resides in the Center for Biology and Society at Arizona State University (ASU). Funded by several grants from the National Science Foundation, the Project has so far brought together an international network of scholars to develop the scholarly content and editorial functions that are producing the online Embryo Project Encyclopedia at <http://embryo.asu.edu>. At the surface the encyclopedia as such can be seen as a more traditional digital project within the HPS community in that it includes a peer-reviewed journal along with collections of archived historical materials such as videos, photos, and manuscripts. However, the Embryo Project team, working with others has also been developing a Digital HPS Collaborative centered at ASU and the Marine Biological Laboratory (MBL) in Woods Hole, Massachusetts.⁴ This consortium of projects is developing new ways of doing scholarly research in HPS based on newly emerging digital tools that help support these endeavors (for further discussion see Maienschein and Laubichler in press).

In 2007 the Embryo Project received its first round of funding from the National Science Foundation, but by that time we had already done a great deal of

⁴ see also <http://www.digitalhps.org>

preparation. In partnership with the ASU and MBL libraries, we decided to adopt the Fedora Commons repository software⁵ and began working with programmers at ASU, the MBL, and the Max Planck Institute for the History of Science (MPI) in Berlin to establish basic ground rules for the data-basing and informatics parts of the project. We soon learned that such efforts are best undertaken within a larger collaborative setting as individual projects, if they want to succeed, grow and persist, need to follow common standards and practices. We soon found that for HPS no community comparable to Bioinformatics existed and that there were a variety of different projects, all more or less trying to reinvent the wheel. This is one reason we are dedicated to work in a collaborative way, since this is the only way to assure interoperability, divide necessary work on cybertechnology, and be able to share limited resources efficiently. We discuss some of our recent technological advances below. The important point here is that we are still building on the Fedora standard and related technologies that many libraries use since we want to be integrated with library archiving, collecting, and federated search functions.

On the editorial side, we worked with scholarly editing experts to establish an ISSN number as a peer-reviewed journal for our new publications. It is extremely important to do this at the very beginning, and not doing so is a mistake many projects have made. We want to be open-access and online, but also a trusted information source that is peer-reviewed and respected at the highest scholarly levels.

We then had to decide what the scholarly content of the Embryo Project would be. That is the main focus of this section. Our dream, also quite insanely naïve but still inspiring to us, was to include everything anybody has said about research related to embryos forever – from Aristotle till tomorrow, as we put it. While the focus would remain on the science of embryo research, we also wanted to contextualize this rich history and include the many layers of cultural, legal, bioethical, policy, and other knowledge relevant to understanding embryos in context as well.

The first step was to decide that although we welcome contributions of all sorts, we would initially provide a framework by looking at several historical episodes that each brought significant changes in theory and practice of studying embryonic development. These included the period of the Hypothetical Embryo, dominated by Aristotle and Aristotelianism. In this period the embryo remained largely a matter of theory in part because of limited technologies and also because of limited reason to ask questions and make observations. Microscopes and renewed enthusiasm for observing and describing that came with the 17th century brought the Observed Embryo, though observations remained on non-human embryos and fetuses. The Biological Embryo period began at the end of the 19th century, with the rise of experimentation and observation inside the developing

⁵ <http://www.fedora-commons.org>

organism. Experimental tools for studying patterns, processes, differentiation, morphogenesis, and other cellular observations brought tremendous advances in the detailed understanding of what embryos are and how they change over time. The Inherited Embryo, starting in the 1950s and '60s, emphasized developmental genetics. Then in 1978, the human embryo became visible from the beginning

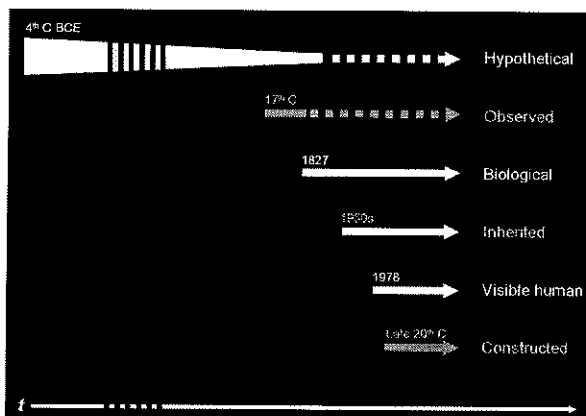


Fig. 1 Episodes in the History of Embryo Research. Courtesy of Jason Scott Robert.

stages with in vitro fertilization. This Visible Human Embryo and its presentations and representations brought considerable public attention. Finally, cloning, stem cell research, and genetic engineering have changed the concept of what an embryo is since researchers can produce the Constructed Embryo. Another period is emerging as well, that of the Computed Embryo, but this episode remains in early stages. These stages are captured below and discussed further in (Maienschein and Robert, in press).

The Project is not constrained by these episodes, but they provide a general framework for thinking about how all the individual pieces fit within an historical framework. Intellectually, what is important about the network of researchers and the approach we have adopted together is that the knowledge produced within the Embryo Project goes far beyond the expertise of any one contributor. While some of us are historians of science, others lawyers or bioethicists or philosophers or historians of technology or other types of specialists, the core Encyclopedia brings all the approaches together. A user of the project therefore does not have to be an expert in legal scholarship to have access to information about laws, information that is linked through the relational database to other articles about the science. The Fedora repository and its store of relationships links ideas that we as scholars, necessarily limited by our own specific training, cannot access without such help.

So far, the Encyclopedia has been working with three main types of entries, but we will be adding more and are already experimenting with some types, as we discuss below. The first of the existing types of entries is photographs and videos.

These existed at the MBL library, and we have digitized them and developed standards for metadata through the libraries and have developed relevant ontologies through our IT team members. To date, we have included more than 1000 photographs of biologists and a few videos, and we will be adding many more related to other areas of science. Since the Encyclopedia is Open Access, working with Creative Commons licensing agreements⁶, we have added considerably to research materials available for scholars and the public with these images. But so far, we have worked with fairly simple collections to get all the details worked out smoothly. We will be adding more kinds of collections soon.

The second type of entry is the short derivative descriptive article, written largely by undergraduate students in a writers workshop seminar setting. These are all heavily and rigorously peer-reviewed before being submitted to the formal editing and publishing process. The products are therefore authored short articles.

Finally, the third type of product is the new scholarly papers by our contributors. The very first of these is the Ph.D. dissertation of Dr. Mary Sunderland (Sunderland 2008), on the history of regeneration research. This was added as a “proof of principle” in order to work out the details of how to digitize such a scholarly article in intelligent and standards-based ways. We have a line-up of other articles to add, but the informatics team is still working out the best way to mark up relationships in long articles (see below for a discussion on relationship mark-up).

We are working with librarians and informatics experts to add a fourth kind of “found object.” These are the publications already out there, and there are many different search approaches for finding them. What we seek is a federated search that will lead the user easily to the kinds of publications the user needs. That takes a lot of sorting of the potentially millions of “hits” for a search, and librarians and informatics experts are working with many different experimental approaches to make searches more effective. We are close to settling on a standard, but are still working on this. When we have effective searches in place, we can add well-chosen short descriptive articles that will find dozens to hundreds of others, including linking to references where those are available online. As a result, what we have is an emerging powerful scholarly Encyclopedia. Our intended users include researchers and also multiple groups of users - students of all ages and the interested public at large. Each needs a different kind of result from a search, and we are also working on refining the tools for diverse user groups.

In the following sections, we introduce work in progress. This is very exciting in its capacities to go beyond what we can do with any existing scholarly tools. What we have described so far is the way an Encyclopedia like ours can link and bring discoveries as a result of finding relationships. Now we turn to new ways of thinking, with completely new kinds of knowledge as a result. The new results in turn lead to new questions and new ways of working.

⁶ <http://creativecommons.org/>

A Cyberinfrastructure of HPS

As mentioned above, the enormous success of the life sciences over the last decades has been facilitated in no small part by the bioinformatics revolution, which has enabled new ways of managing, annotating, and sharing data. It also increasingly allows for the integration of different kinds of data, such as genomic and pathological data, genomic and environmental data, pathological and population data, genomic and historical data, etc. Fundamentally the computational challenges for HPS are very similar to those in the life sciences, but, as we have learned in the context of the Embryo Project, existing computational solutions need to be adapted and some novel challenges for history and philosophy of science need to be addressed as well. Specifically these challenges include

- ontology development
- natural language processing and semantic web tools
- digital workbench applications and web interfaces.
- database design and management

Of these, ontology development and database design are the most crucial and foundational. Unlike in the life sciences, there are not yet established and accepted ontology standards for HPS or HPS-related fields, partly due to the late adoption of digital technologies into the humanities and partly due to the difficulties of identifying and parsing the subtle interpretations that are indicative of the abstract language used in many humanities writings (Niepert et al. 2007).

We have begun to develop a working ontology for the Embryo Project and several tools that allow us to utilize our ontology to extract meaningful information from digital texts, which is a good start but requires further testing and refinement. Our tools are based on the Open Biomedical Ontology standards⁷ and are based on scripts that mine texts and extract meaningful information and relationships within the context of large data sets. However, the data relevant for historical analyses are generally varied and include different data types (ranging from different styles of literature to images, audio and more recently video). As a result, these scripts need to be further developed and tailored to multiple specific databases. The challenges are substantial and (bio) informatics is a broad and complex field. Nonetheless, the Embryo Project has already proven that applying informatics approaches can also transform HPS projects.

Informatics in biology and the biomedical fields works because all participants adhere to a strict set of standards about data representation and publication, which includes agreed-upon ways to make primary data available for others to use. In this way, large-scale genomics studies that span multiple species can be performed. Standardizing the ways information is captured and stored is integral to the effective management and sharing of knowledge. The world-wide web is itself a standard that allows people to exchange information via the creation of websites

⁷ see <http://www.obofoundry.org/>

that contain all kinds of content. Microsoft's document format has become a standard in many academic and non-academic circles for the creation and sharing of textual information.

If one wants to be involved with the larger community that exchanges information in this way, one must make sure that information is captured and represented in the standard format. Meaningful collaborations that facilitate sharing of knowledge effectively build upon standardized sets of tools and methodologies. The term "cyberinfrastructure" is used to describe the sets of technologies and methodologies that enable data to be acquired, managed, and analyzed effectively so that disparate projects can share information. For example, in the biomedical field there are libraries of ontologies that describe entities and their relations to one another within specific research domains (Smith et al. 2007). The mission of the OBO Foundry specifically states their hope that "ontologies will be fully interoperable, by virtue of a common design philosophy and implementation, thereby enabling scientists and their instruments to communicate with minimum ambiguity. In this way the data generated in the course of biomedical research will form a single, consistent, cumulatively expanding, and algorithmically tractable whole."⁸

As outlined above, the Embryo Project is one such project in the area of science studies that has adopted the principles of biomedical informatics and has taken an open access approach. Based on the advice of the developers and librarians at the ASU library, the MBL and the MPI, the Embryo Project has adopted protocols that ensure that we are following common standards and practices and adopting technologies and platforms that make the Project open and freely and easily accessible. All articles are marked up under the National Library of Medicine's journal article DTD, which ensures long-term sustainability and exchangeability with a broad range of collaborators. Moreover, citations in the Embryo Project are tagged with MODS metadata⁹, a schema developed by the Library of Congress for bibliographic entities. Likewise, images are collected and stored in the JPEG2000 format, which is fast becoming a standard for archiving digital images.

The Embryo Project has adopted a set of practices that rely on W3C standards for creating content and storing relationship information among objects, and has taken great care to annotate textual articles with these relationships in mind. Presently, all articles are marked up in the XHTML format and hand-coded with relationship information that is relevant to the various aspects of embryo research. This relationship information is currently referenced from an informal ontology of categories - People, Places, Organization, Contexts, Awards, Concepts, Law, Ethics, Religion, Technology, Experiments, Organisms, and Literature - and are stored as a RELS-EXT data-stream in the Fedora repository. Articles, images, and

⁸ <http://www.obofoundry.org>

⁹ <http://www.loc.gov/standards/mods/>

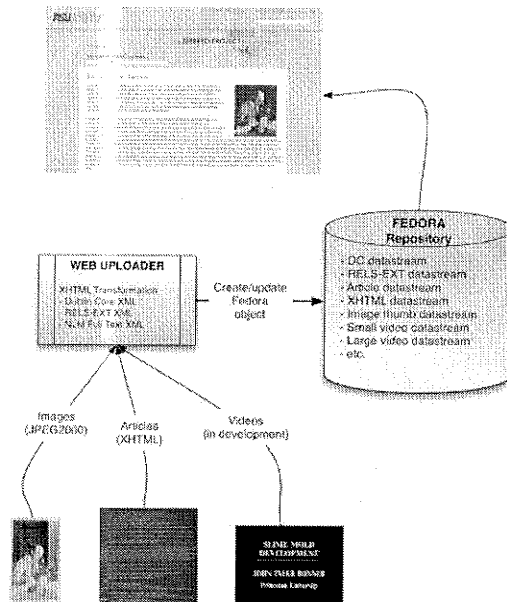


Fig. 2 Embryo Project Workflow for creating, ingesting, and managing objects. Objects are uploaded to the repository via a web interface (currently only images, but will soon be enabled for images and videos). XHTML transformations take the text and extract and create XML representations that are then stored as datastreams in Fedora, which are disseminated via the website.

videos are put into Fedora via a web uploader that the library has developed, which gives control of the ingestion process to the content creators. The information contained within the XHTML document is parsed and stored as separate datastreams in the Fedora repository. The relevant data-streams are then displayed in the encyclopedia via a web browser (see Figure 2).

Until now, technological contributors to the Embryo Project have devoted most of their effort to developing workflows and identifying best practices to address core questions about how to create unique content, edit and publish this content, and manage this content effectively. Phase One of the Embryo Project, then, has been about creating and managing its own content. This, however, is only the beginning.

The next stage of the Embryo Project involves development of an expansive and robust repository of embryo-related information that accesses other repositories and incorporates the vast sources of already-published materials on embryo research. In order to do this, the Embryo Project requires an informatics approach to mining, extracting, and analyzing content from diverse repositories. To accomplish this effectively the following challenges that the Embryo Project faces in its phase two, the semantic web phase, need to be met (these are the same challenges that all other digital projects in these areas of scholarship are also facing):

Ontology development

A formal ontology is critical to any digital project, to organize and structure the relevant terms and their relationships to one another. This ontology will be a necessary component of using semantic web technologies to do text mining and natural language processing of text documents in both the Embryo Project repository as well as other repositories. We have begun the difficult process of ontology development utilizing common tools in the informatics community (OWL, Protégé), but ontology development is an ongoing part of all digital projects (Noy and McGuinness 2001; Horridge et al 2007). One approach that we have been taken is to utilize multiple Obo-Foundry based ontologies in order to facilitate data mining and annotation (see below).

Text mining and natural language processing (NLP)

In order to extract relevant information from text, we must be able to develop tools that recognize not only exact terms matched against some ontology, but also perform natural language processing by recognizing words based on structures such as parts of speech (Riloff 1999). This will enable us to access large amounts of textual information and analyze its content computationally (see Hamed and Sarkar forthcoming).

Working with other repositories and databases

In order to increase both the size and range of objects in our repository, we need the capability of accessing other databases to cull relevant information and populate our repository with either the content or references to the content. In the larger HPS community, there are no federated databases in the way that PubMed aggregates text sources in the biomedical field. Because there are so many separate databases, many of which are specific to a journal or publisher, many different strategies for accessing the content are required.

RDF creation, storage, and use

As part of the Embryo Project workflow of writing original articles, we are careful to include relationship information within the source of that article. These relationships are stored as a RELS-EXT datastream that contains the RDF metadata of all the objects and their relationships to one another (Brickey and Guha 2004). We have not yet determined how to utilize this rich source of information, but its use is critical if we are to leverage the power of the semantic web. If the sharing of data, rather than documents, is to be accomplished, we need to develop standards among HPS digital projects so that RDF representations are captured and made available to other projects (Allemang and Hendler 2006). Tools

that facilitate the exchange of RDF information as well as its analysis within the scope of the project are also important.

Each of the above problems is not unique to the Embryo Project, but for all fields of research. HPS research currently does not have the ability to leverage the informatics tools being used in the life sciences to solve these problems. Right now, we are working with our partners in Digital HPS to take initial steps in that direction to address each of the issues and understand the kinds of technologies and expertise necessary to move forward. In general, two main problems need to be addressed: (1) how to analyze large amounts of textual information computationally and (2) how to share this data with others.

Currently, all our articles are marked up and annotated by hand. While this has the advantage of being accurate, it does not scale when hundreds and thousands of articles need to be annotated. NLP tools, like OBO-Annotator and Ontotator (Hamed and Sarkar, forthcoming), are able to analyze text and, run against one or more ontologies, extract relevant terms in these ontologies.

We also have developed an extension of the OBO-Annotator tool, which we have called Vagon, that takes a text, displays any instances of terms in an ontology, and provides a user interface that allows researchers to identify relationships based on a standardized relationship ontology (see Figure 3 and 4). This information is then stored in a relational database that can be searched. This kind of NLP tool not only extracts terms in the ontology, but performs NLP tasks such as sentence tokenization and part-of-speech tagging. The data can be stored in a relational database and transformed into RDF triples that are then deposited into a triple store, which is a large repository of statements with the subject-predicate-object schema. Alternatively, these triples can be extracted directly from the text mining process. This triple store is the repository of all data of all the relevant content in the articles that were analyzed. It can be queried utilizing languages such as Prolog and SPARQL, and a large, federated triple store can also be created by incorporating triples that have been extracted in other projects.

The Encyclopedia of Life¹⁰, for example, is creating this kind of federated triple store that will contain little bits of information about all species on earth. Researchers on other projects are encouraged to provide their triples to the Encyclopedia of Life, thus increasing both the amount and variety of information on species culled from projects with different interests. In this way, data about a species' habitat and range can "live" alongside data about a species' embryology. This promises not only a single pool of all information on species, which is the main goal of the Encyclopedia of Life, but holds the added promise of creating new knowledge because queries to the triple store will reveal relationships that were not detectable by any single project or database. The potential of sharing data among projects is now realized, with the further goal of inferring new knowledge

¹⁰ <http://www.els.org>

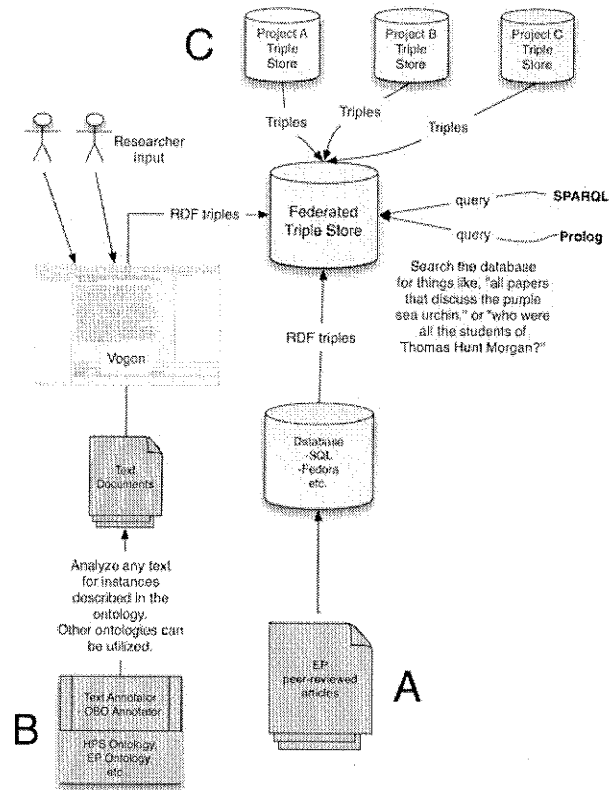


Fig. 3 Ontology Workflow. In our scenario, a federated triple store would incorporate triples produced in a number of ways. (A) Embryo Project peer-reviewed articles are marked up utilizing a specific ontology. Relationship information is rigorously checked and hand-curated so that each article that is incorporated into the Encyclopedia also provides a rich set of RDF information that is stored in the triple store. (B) Various ontologies are utilized to analyze text documents using an application like OBO-Annotator (or Ontotator). The results of the analysis are passed on to Vogon, which highlights terms that OBO-Annotator found and allows end-users to edit relationship information that is then stored as RDF triples in the triple store. (C) Other digital projects produce their own triples and share this information with each other. This federated triple store can then be queried utilizing query languages such as SPARQL and Prolog. The richer the store of information contained in the triples, the more interesting and novel the results may be, ideally leading to discoveries that would not otherwise be possible.

from this pool of collective data (Allemang and Hendler 2006). Figure 3 outlines a simplified view of how this informatics approach works.

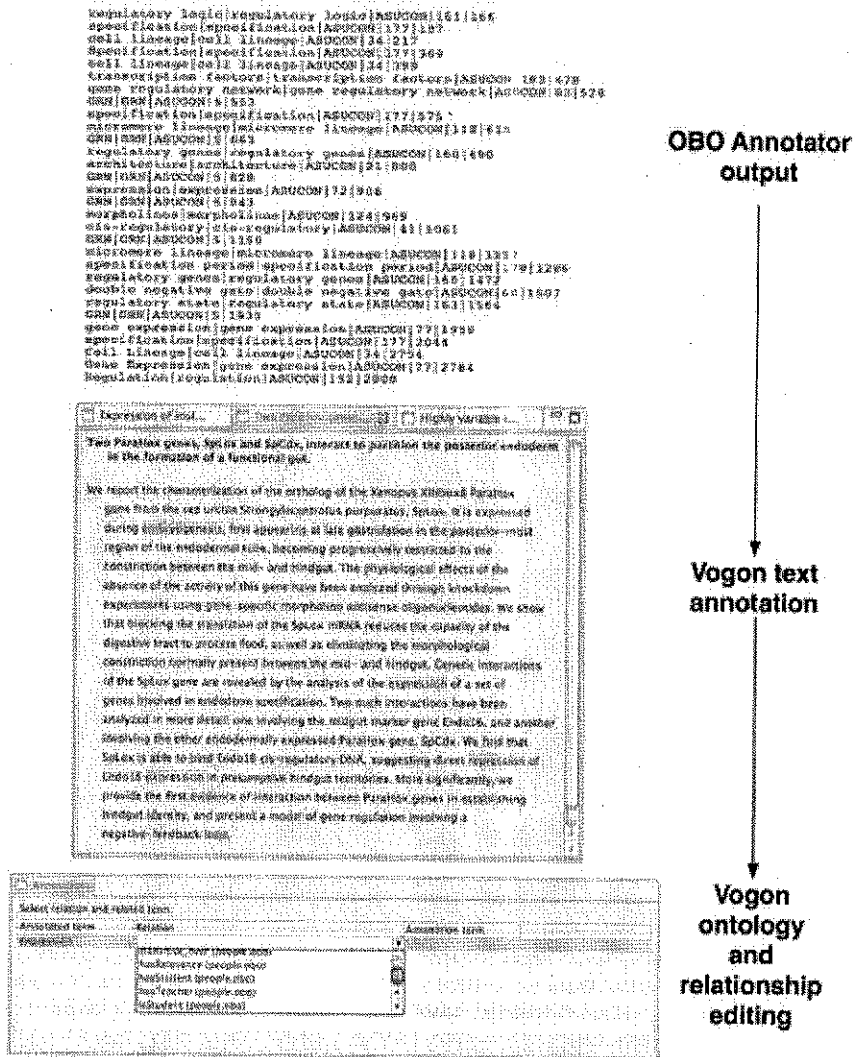


Fig. 4 OBO-Vogon Tool. OBO-Annotator analyzes text documents and outputs found terms in the ontology and their locations within the text. Vogon takes this output and highlights these terms. Based on another relationship ontology, the user is then able to specify the relationship of the found term to other terms in the document. These “triples” relating two terms to one another are then passed on as RDF statements to the triple store.

In addition to the database and annotation tools, the Embryo Project has also adapted and further developed visualization and analysis tools that enable and support the interpretation of our data for a wide range of users and scholarly interests. These tools include a test-case front-end web application that highlights concepts by matching their locations within a text. Thus, rather than searching for a specific string pattern and highlighting that string, the application highlights the location of the concept. These locations are important for future development of the semantic web aspects of the application, since one can then start inferring meaning within the text by connecting the relative proximities of concepts (and in

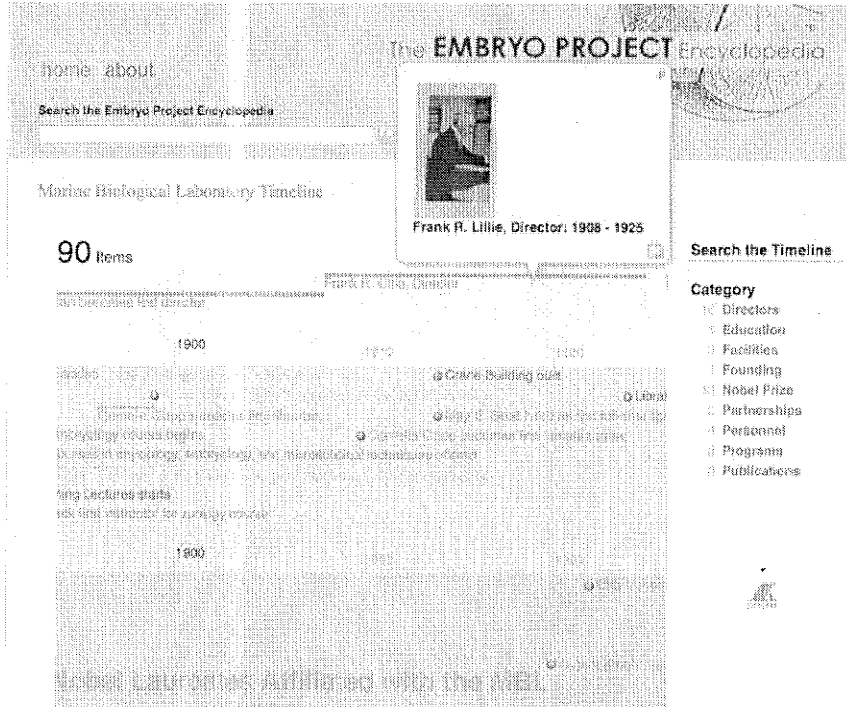


Fig. 5 Timeline Tool. Timelines are one technology that allows rich textual historical information to be visualized in a clean interface yet still allow for user interaction with the technology. Rather than providing a static set of information, each datum within the timeline contains links to other objects (articles, videos, images) within the repository or links to other relevant websites. In the timeline for the Marine Biological Laboratory, for example, we are able to aggregate all the directors of the MBL and juxtapose this information against various important events in the history of the MBL as well as note the various Nobel Laureates that were affiliated with the MBL. The technology used for this timeline is adapted from MIT's SIMILE framework, an open source project that rests on standards that any web browser can read and understand. We have developed workflows that make it easy for a student, researcher, or anyone interested in the history of a particular subject to build these timelines.

the future, organisms, relationships, experiments, people, and institutions). The concepts within a text are therefore unique because they occupy unique locations within the text, so "embryogenesis" in the beginning of an abstract is distinguished from "embryogenesis" later on.

This also allows us to take advantage of several available tools that highlight networks between concepts and allow us to dynamically refocus the display. For example, starting with one concept, such as "cis-regulatory logic" in a text, one can, by simply clicking on it in the text displayed in the application, get all additional papers that mention this concept, all authors associated with that concept, etc. Or, one can refocus on one author and immediately get all co-authors or all concepts that this author is connected with, etc. With such dynamic applications it is thus possible to test multiple scholarly hypotheses at an unprecedented scale and to reveal patterns that are otherwise difficult to detect. Similarly, timeline applications also allow the user to display a lot of information simultaneously based on a temporal ordering principle. As Figure 5 shows, based on one of the focal projects within the EP—the history of the Marine Biological Laboratory - such tools also enable us to display events in a variety of different categories, thus allowing to highlight events of interest, such as a Nobel Prize, within a variety of contexts, such as the availability of new technologies or facilities.

Conclusion — Challenges and Possibilities of Digital HPS

Therefore the development of a cyberinfrastructure for the science studies community will have to follow the same distributed approach it will ultimately enable. But, even though the only way to make substantial and sustainable progress is through an organized community effort, something the Embryo Project team has been organizing with the support of NSF, this approach also requires that we change the culture in HPS to include both a scholarly and an informatics perspective and to focus more on interdisciplinary teams of informatics experts and scholars working on these challenges. One additional benefit of an informatics approach to HPS is that it will also allow for a much closer integration of historical and philosophical scholarship into current science. HPS used to be closely aligned with scientific endeavors and disciplines, but has unfortunately lost some of these ties as it developed its own disciplinary identity. Rebuilding these bridges is essential for fields that are, by their very nature, interdisciplinary. And in the context of 21st century science this means that HPS needs to be integrated utilizing the new informatics tools as those have become the way scientists conduct business.

The biggest challenge to success is that the HPS community needs to take advantage of existing informatics tools and approaches. We do not need to start over and reinvent what has already been widely tested and proven elsewhere. Yet we do need to develop tools to capture the unique aspects of HPS research,

including especially the importance of time and changes over time. Unfortunately, existing funding agencies are focused on supporting new computing technologies rather than on developing effective applications in defined communities. We recognize the need to carry out much of this work ourselves, collectively and collaboratively, sharing our tools and approaches as well as our scholarly discoveries. HPS scholarship must embrace traditional research and also development of new ways of working, which in turn leads to new discoveries. We are encouraged by the support for such types of development in various places, such as the Max Planck Institute for the History of Science, The Marine Biological Laboratory, Indiana University, Caltech, ASU and many museum and library based projects. What is still lagging behind is an improved coordination between these different projects.

We therefore call for cooperative and open development of digital HPS tools and collaborative working groups through Digital HPS.

Acknowledgments. This project has benefited tremendously from grant support from the National Science Foundation (SES 0623176, 0914069 to JM and MDL and 0645729 to MDL) and Arizona State University, and from the support and guidance of our institutional partners at the Max Planck Institute for the History of Science in Berlin and the Marine Biological Laboratory-Woods Hole Oceanographic Institution Library. Felicity Snyder and Andrew Hamilton have provided far more in the way of intellectual leadership than can be indicated on the website, as have Marie Glitz, Mary Sunderland, and Cera Lawrence. Julia Damerow has developed the Vagon tool. The ASU Libraries Informatics and Cyberinfrastructure Services team, led by librarians John Howard and Philip Konomos, have played indispensable intellectual and implementation roles. Jacob Sahertian and the SOLS Vis Lab have also provided much website support. MBL-WHOI Library Director Cathy Norton and MBL Director Gary Borisy have provided tremendous support and encouragement, as has the MBL technical.

References:

- Alon, U. (2007). *An Introduction to Systems Biology. Design Principles of Biological Circuits*. CRC Press, Boca Raton, FL.
- Allemang, D., Hendler, J. (2006) *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. Morgan Kaufmann, San Francisco.
- Bonner, J.T. (2008) Bonner Workshop.
<http://embryo.asu.edu/about/projects.php>.
- Brickley, D., Guha, R.V. (2004) *RDF Vocabulary Description Language 1.0: RDF Schema*, W3C Recommendation. W3C: <http://www.w3.org/TR/PR-rdf-schema>.

- Cantara, L. (2004) Digital libraries in the humanities: Building a cyberinfrastructure for the humanities. *OCLC Systems & Services: International Digital Library Perspectives* 20, pp.167-169.
- Creative Commons (2009) <http://creativecommons.org>.
- Digital HPS Collaborative (2009) <http://www.digitalhps.org/>.
- The Embryo Project (2009) <http://embryo.asu.edu/>.
- The Encyclopedia of Life (2009) <http://www.eol.org>.
- Fedora Commons (2009) <http://www.fedora-commons.org/>.
- Hamed, A.A., Sarkar, I.N. (Forthcoming). OBO-Annotator: An OBO Document Analyzer Application. *Bioinformatics*.
- Horridge, M., Jupp, S., Moulton, G., Rector, A., Stevens, R., Wroe, C. (2007) *A Practical Guide to Building OWL Ontologies Using Protégé 4 and CO-ODE Tools Edition 1.1*. The University of Manchester. <http://www.co-ode.org/resources/tutorials/ProtegeOWLTutorial-p4.0.pdf>.
- Kitano, H. (2001) *Foundations of Systems Biology*. MIT Press, Cambridge.
- Lesk, A. (2008). *Introduction to Bioinformatics*, 3rd edition. Oxford University Press, Oxford.
- The Library of Congress (2009) *Metadata Objects Description Schema*. <http://www.loc.gov/standards/mods/>.
- Maienschein, J., Robert, J.S. (in press) What is an embryo and how do we know? In: *The Healthy Embryo* (Nisker, J. et al.; eds.) Cambridge University Press, Cambridge.
- Maienschein, J., Laubichler, M. (in press) The Embryo Project: An integrated approach to history, practices, and social contexts of embryo research. *Journal of the History of Biology*.
- Martin, S. (2007) Digital Scholarship and Cyberinfrastructure in the Humanities: Lessons from the Text Creation Partnership. *Journal of Electronic Publishing* 10. <http://dx.doi.org/10.3998/3336451.0010.105>.
- National Center for Biotechnology Information (2009) <http://www.ncbi.nlm.nih.gov/>
- Niepert, M., Buckner, C., Allen, C. (2007) A Dynamic Ontology for a Dynamic Reference Work. *JCDL 2007 Proceedings*, pp. 288-297. <http://doi.acm.org/10.1145/1255175.1255230>.
- Noy, N.F., McGuinness, D.L. (2001) *Ontology Development 101: A Guide to Creating Your First Ontology*. Stanford Knowledge Systems Laboratory:

<http://www-ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf>.

The Open Biomedical Ontologies (2009) www.obofoundry.org.

Ouzounis, C.A., Valencia, A. (2003) Early bioinformatics: the birth of a discipline - a personal view. *Bioinformatics* 19, pp. 2176-2190.

Ouzounis, C.A. (2002) Bioinformatics and the Theoretical Foundations of Molecular Biology. *Bioinformatics* 18, pp. 377-378.

Renear, A.H. and Palmer, C.L. (2009) Strategic Reading, Ontologies, and the Future of Scientific Publishing. *Science* 325: 828-832.

Riloff, E. (1999) Information Extraction as a Stepping Stone toward Story Understanding. In: *Understanding Language Understanding: Computational Models of Reading* (Ram, A., Moorman, K.; eds.) The MIT Press, Cambridge: pp. 435-460.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L., Eilbeck, K., Ireland, A., Mungall, C. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 25, pp. 1251-1255.

Sunderland, M.E. (2008) Studying Development: The Value of Diversity, Theory, and Synthesis. http://embryo.asu.edu/articles/article_sunderland.php.

Address for Correspondence:

Prof. Dr. Manfred D. Laubichler
School of Life Sciences
Arizona State University
PO Box 4501
Tempe, AZ 85287-4501
USA
Manfred.Laubichler@asu.edu