The Archimedes Project: Realizing the Vision of an Open Digital Research Library for the Study of Long-Term Developments in the History of Mechanics

U. Schoepflin
Director of Library
Max Planck Institute for the History of Science
Wilhelmstr. 44
D-10117 Berlin
Germany
Tel.: ++49-30-22667-190
Fax: ++49-30-22667-299
email: schoepfl@mpiwg-berlin.mpg.de
http://www.mpiwg-berlin.mpg.de/ENGLHOME.HTM

Abstract

The Archimedes Project will create a testbed for developing and exploring model interactive environments for the history of mechanics. It will also serve as a proof-of-concept project for open digital libraries for topics in the history of science designed to integrate research and knowledge dissemination in new ways.

The project is funded by the Digital Libraries Initiative Phase 2 program of the National Science Foundation (NSF) and of the Deutsche Forschungsgemeinschaft (DFG) and is a joint endeavor of the Classics Department at Harvard University, the Max Planck Institute for the History of Science (MPIWG) in Berlin, the English Department at the University of Missouri at Kansas City, and the Perseus Project at Tufts University.

Numerous treatises on mechanics as well as other forms of documentation of mechanical knowledge and practices constitute the project corpus. Ongoing research at the MPIWG on the long-term development of mental models of mechanical thinking and their manifestation in technical terminologies, inferences of practitioners, engineers, and scientists plays an important role in the testbed design. The testbed also requires a powerful, linguistically based information technology for handling the variety of languages occurring in the source materials. Source documents are being prepared with tools such as automatic morphological analysis of Latin, Greek and Italian, and semantic linking of sources to general and technical, modern and historical dictionaries and reference works.

Proceedings of the 5th Russian Conference on Digital Libraries RCDL2003, St.-Petersburg, Russia, 2003

1 The institutional context of the Archimedes Project

The Archimedes Project [1, 2] deals with the entire tradition of mechanical knowledge that is so closely associated with the name of Archimedes and his achievements. The Archimedes Project is a proof-of-concept endeavor for an open digital research library in the history of science that proposes to combine research and dissemination in a new way. It closely integrates scholarly work and technical developments.

The project would not have been possible without the tight collaboration of American and European partners who have not only invested previous experience in this project, but who also share a commitment to open access without which the vision of an open digital research library would be impossible to accomplish. The Archimedes Project is a joint endeavor of

- the Max Planck Institute for the History of Science (MPIWG) in Berlin,
- the Classics Department at Harvard University,
- the English Department at the University of Missouri at Kansas City, and
- the Perseus Project at Tufts University .

The project is being executed within a wider network of scholarly cooperation including the Programme International de Coopération Scientifique (PICS). It is supported by the Deutsche Forschungsgemeinschaft (DFG) and the National Science Foundation (NSF) of the United States in the framework of the NSF-DFG International Digital Library Initiative.

The German principal investigators of the project are:

- Jürgen Renn (MPIWG)
- Peter Damerow (MPIWG)
- Urs Schoepflin (MPIWG).

The US principal investigators of the project are:

- Mark Schiefsky (Harvard University)
- Jeffrey Rydberg-Cox (Univ. of Missouri)
- Gregory Crane (Tufts University).

The project staff includes Brian Fuchs (MPIWG), Malcolm D. Hyman (Harvard), Elaheh Kheirandish (Harvard), Peter McLaughlin (MPIWG), Marcus Popplow (MPIWG), Markus Schnöpf (MPIWG).

2 The scholarly context of the Archimedes Project

The Archimedes Project is the digital library component of a major research project of the Max Planck Institute for the History of Science dealing with mental models in the history of mechanics. In the context of this project, research is focused on questions concerning the long-term development of mechanics such as the following:

- How could atomism have been invented in Greek antiquity long before the first experiments revealed the real existence of atoms?
- Why did Aristotelian natural philosophy dominate scientific thinking for more than 2000 years in spite of obviously fallacious assumptions such as that of the speed of fall is proportional to the weight of the falling body?
- How was the complicated technology of architecture and machine construction possible without the scientific base that we associate today with such achievements?
- How was it possible to overcome the concepts of classical mechanics by a radically new theory of space and time as it was formulated by Einstein long before physical and astronomical evidence suggesting such a theory was available?

Such questions can hardly find satisfactory answers in the study of the universal structure of the science, as is customary in philosophy, nor in studies focused on specific disciplines during specific historical periods based on a small selection of sources as is customary in historiography.

The Max Planck Institute for the History of Science has therefore undertaken a study of the long-term development of mechanical knowledge which encompasses the vast historical period from the third millennium B.C. until the decline of classical mechanics and currently focuses on scientific as well as practical knowledge from contrasting cultures such as ancient Mesopotamia, classical Greece, ancient China, the world of the Arabic and Latin Middle Ages, and the early modern period of the European engineer scientists.

In this vast period mechanical knowledge has completely changed its character several times. Mechanics has its origins in everyday experience built up in roughly the same way by individuals under extremely different cultural settings such as shipbuilding in the South Pacific or the use of ballistic weapons ranging from blast pipe technologies to throwing boomerangs. For a long time even highly-developed technologies of this kind were not accompanied by explicitly documented mechanical knowledge of the kind familiar from later mechanical treatises.

When the first treatises on mechanics emerged, their practical relevance was negligible. Rather, they profited more from existing mechanical knowledge when formulating theoretical concepts with a potentially unlimited range of application than they contributed to the structuring and advancement of the practical knowledge embodied in professional activities. This relation underwent change only during the Renaissance, when challenging new tasks required that the entire realm of mechanical technologies be drawn on, which led to a realization of the implicit potential of the theoretical resources accumulated since antiquity. This realization was the beginning of an age in which the growing integration of theoretical and practical knowledge accompanied the advancement of technology from a marginal phenomenon to a central element of social productivity in the economy of modern capitalist society. The changing role of theoretical mechanical knowledge was foreshadowed in the realm of ideas where the so-called mechanistic world view emerged and then developed into an almost universal and intuitively self-evident doctrine. This mechanical world view was abandoned only when the rich results of science, to which it had contributed, could no longer be fitted into its framework and when twentieth-centry physics made its seemingly timeless conceptual foundations obsolete.

The study of such long-term processes of the development of scientific knowledge requires not only a general theoretical framework of a sort unlikely to emerge from specialized historiography or from rational speculative syntheses. Such a study obviously also hinges on the analysis of an enormous corpus of sources that serve as its empirical validation. These sources have to cover

mechanical knowledge far beyond what is usually taken into account by a history of science that focuses on individual discoveries. What the project at the MPIWG is attempting to reconstruct for the different historical periods covered by the project are the following dimensions of knowledge:

- theoretical scientific knowledge,
- the practical knowledge implicit in the productive use of technology which is a common precondition of both science and technology, and
- the universal knowledge embodied in cultural practices similar in all human cultures.

This may explain the two main aspects of the research which forms the foundation of the Archimedes Project :

- the application of methods and techniques of cognitive science in order to capture the structures that organize the interaction of these three types of knowledge,
- the use of computer-assisted analysis that enables us to extract information relevant to the reconstruction of such structures on the basis of historical sources.

3 The production line of the Archimedes Project

To achieve these goals, a variety of qualitatively different sources are being digitized for a digital research library which will represent the knowledge base required to study the long-term development of mechanical thinking. These sources, which are made freely available to the research community at large as part of the Archimedes Project, include theoretical texts and documents of practical knowledge from different historical contexts and in different languages (mainly Greek, Latin, Arabic, Italian, and English); images of artefacts of mechanical technology; and also video documentation of activities guided by intuitive physical knowledge as well as of still extant traditional techniques involving mechanical knowledge.

Accordingly, the production line of the Archimedes Project comprises three major dimensions:

- the digitization and standardization of sources,
- the extraction of organizing structures of shared knowledge as well as those characteristic of innovative processes,
- the creation of a content-based access system to the source repository.

For all three dimensions, electronic tools are being developed which allow for the integration of technical and scholarly competence in analyzing the sources. Given the nature of the Archimedes Project as the backbone of a wide-ranging interdisciplinary research project, this integration of different competencies should, however, not remain limited to the circle of the institutions who bear primary responsibility. The project aims in fact at becoming a growing international digital research library that integrates more and more research institutions, archives, and libraries, offering not only direct and open access on the Internet to primary materials usually restricted to a small group of scholars but also a platform for collaborative research on these sources.

4 The technology applied in the Archimedes Project

The way in which the Archimedes Project deals with the digitization of historical sources and their analysis is determined by the necessary interplay between technical and scholarly work. This interplay is reflected in the development and use of some key instruments of the project. Indeed, the first phase of the project was, apart from the continuation of data entry, mainly dedicated to the development of these instruments and their implementation in the project's production line.

The production line follows a number of well-defined steps:

- text selection and the data entry of sources,
- automatic minimal XML tagging,
- interactive tagging of formal source structures such as chapter divisions,
- automatic generation of metadata such as the morphological analysis of words and the establishment of links to an integrated system of sources and metadata, such as dictionaries and bibliographic references,
- interactive creation of scholarly metadata within the content-based access system.

All of these steps require tools, which are continuously being refined in tandem with the improvement of the production line. The aim of the project is to make all the tools openly available as soon as they have reached an appropriate state of development.

Currently the following tools are freely accessible and – if appropriate – can be downloaded from the Archimedes workbenches in Berlin [4] and in Harvard [5] respectively:

- SGML Parser (parsing and analyzing an SGML document)
- Filemaker Morphology Tool (Morphological analysis for Filemaker)
- Filemaker XML Input Tool (Convert XML to Filemaker input)

- Arboreal (content-based XML browser and annotation environment)
- Donatus (morphological analysis for XML texts)
- Pollux (direct access to dictionaries)
- Ficus (automatic insertion of page image and figure attributes)

Three key instruments will briefly be described:

- "Digilib," an image server for which pilot installations with some 10.000 images exist on two servers, one in Germany, the other in Switzerland,
- an image and text display environment for which a test implementation exists on the internal Archimedes server in Berlin,
- "Arboreal," a working environment for content analysis allowing the production of scholarly metadata--although this is still under development, a release with basic functions can be down-loaded from the Harvard Archimedes server.

In addition to these three key instruments, a content-based access system allowing collaborative scholarly work to be performed on the web and to turn the results of such work into navigation devices for the Archimedes sources is currently under design.

4.1 The image server "Digilib"

"Digilib" is an image server with annotation facilities. It has been developed in close cooperation with the University of Bern where Gerd Grasshoff, a former member of the research group at the Max Planck Institute for the History of Science, now holds the chair for the philosophy of science.

The image server enables scholars to collaborate via the web on a distributed collection of images whose high-speed transmission to a local site is made possible by pre-scaling the images before they are transferred. Apart from consulting a scaled image, the user has the possibility of zooming in on the high resolution image stored on the server and of placing up to eight marks per page which may be used as references for linking commentaries to locations on the image via URLs provided by the image server. The image server furthermore generates on-the-fly thumbnails of the images which assist orientation in lengthy texts.

4.2 The basic internet display environment

While the image server allows sources to be dealt with as soon as digital reproductions are available, that is, in an early phase of the production line, the Archimedes basic display environment makes it possible to combine the display of digital facsimiles with the transcribed Archimedes texts, which in turn are enhanced by language technology. Even in its provisional form it allows for browsing with the help of thumbnails, page images, and text retrieval, and also allows for connection to high-resolution images of a given page.

The language tools, which are based on an implementation and further development of the Perseus technology hinge on the idea of combining the morphological analysis of a word with a link to one or several freely available dictionaries. Following up on earlier work of the Perseus Project on ancient Greek and Latin, and in close collaboration with the Perseus Project, the Archimedes Project is building analogous environments for Italian (which is available in a provisional form), as well as Arabic, Dutch, English, and German (which are under development).

The crucial point of this technology is the choice and the unrestricted availability of appropriate dictionaries. At present, a modern English-Italian dictionary has been made available, as well as two historical dictionaries from the early modern period, one for Italian-English, the other for Latin-English. A major Arabic-English dictionary is currently being digitized.

In addition to the access to dictionaries that are integrated in the language technology of the standard display environment allowing for lemmatized corpus searches, Archimedes offers direct open Internet access to number of relevant dictionaries and encyclopedias in a special section of the Archimedes site where they can be accessed at headword level. Thus the corpus of source texts for the study of the history of mechanics is enhanced by historic standard reference works. The dictionaries available with the headword look-up tool currently include:

- H. Anthony Salmoné: An Advanced Learner's Arabic-English Dictionary,1889
- Florio, John: Dictionarie of the Italian and English Tounges, 1611
- Thomas Cooper: Thesaurus Linguae Romanae et Britannicae, 1584
- Lewis and Short: Latin Lexicon
- Liddel and Scott: Greek Lexicon
- Comelati, Guglielmo, Davenport, John: A New Dictionary of the Italian and English Languages, 1854
- Hutton, Charles: A Mathematical and Philosphical Dictionary, 1795
- Gehler, J. S. T.: Physicalisches Wörterbuch, 1787
- Gehler, J. S. T.: Physicalisches Wörterbuch, 1798

4.3 The XML tool "Arboreal"

True content-based access requires a tight integration of scholarly work and technology. Based on research in the history of mechanical knowledge at the Max Planck Institute for the History of Science, the structures documenting this knowledge in the historical sources can be systematically identified with the help of formal and linguistic clues in these sources. Among these structures are concepts expressed in technical terminology, mental models shaping, for instance, ideas about how a balance works - typically reflected by aggregations of technical terms - as well as larger chunks of content that are more or less coherently transmitted in history and represented by entire texts or larger parts of texts dealing, for instance, with the so-called "simple machines." The Archimedes Project has therefore been developing a working environment for content analysis which not only helps to identify these structures but also to generate the scholarly metadata, harvesting the results of such an analysis in the form of structured annotations to an XML source text. The idea of the "Arboreal" software, conceived specifically for this purpose, is to create an editor supporting the identification of technical terms, mental models, and chunks of content but which also allows for a comparison between different text editions and translations.

An example may elucidate this idea. One of the key sources in the Archimedes collection is a treatise on mechanics published in 1577 in Latin by Galileo Galilei's patron Guidobaldo del Monte. This text represents the first comprehensive early modern treatise on mechanics synthesizing both contemporary and ancient knowledge. Four years later Guidobaldo published an Italian translation of this text prepared by Pigafetta.

Assume now that the XML file of Guidobaldo del Monte's original Latin version has been loaded into "Arboreal" as a "master text" together with Pigafetta's Italian translation and with a modern English translation as "slave texts," matched to the Latin version by way of the XML structure. Chunks of these three texts automatically appear - rendered via Unicode - in separate windows of Arboreal as soon as they are selected in the main window of the XML master text. In order to facilitate browsing, Arboreal comprises links to figures in the text and also allows page images to be loaded.

For an illustration of the functionalities that "Arboreal" provides for analyzing the text, take the example of a study on the role of Aristotelean natural philosophy in early modern mechanics as is represented by Guidobaldo's text. Did Guidobaldo still make use of the Aristotelean distinction between natural and violent motion? Or does a concept such as that of a "neutral motion" - in some sense a predecessor of the classical notion of inertia - already make its appearance in this central text of pre-classical mechanics?

Using the sophisticated search facilities of Arboreal including morphological forms and regular expressions, candidates for the term "neutral motion" can be easily searched and displayed in a separate window as well as highlighted in the master text window. Working in this way on the terminology for motion, a term list can be created and instances can be attached to the list. Such lists can be saved and reloaded, and even used for seeding the terms to other texts.

The particular list which can easily be generated in this way makes it immediately evident that Guidobaldo still used the traditional Aristotelean distinctions, but also that he introduced the term "neutral motion" which played such a prominent role in the later work of Galileo. Each of the terms in this list is linked to instances in the master text that have been identified as pertaining to the relevant term. The instances already identified can now actually be used to navigate further through the text, thus providing a powerful example of content-based access.

4 Outlook

The Archimedes Project began 1999 as a joint initiative of the Max Planck Institute for the History of Science (MPIWG) in Berlin, the Classics Department at Harvard University, the English Department at the University of Missouri at Kansas City, and the Perseus Project at Tufts University. The first project phase coming to an end, what could be achieved of the envisioned open digital research library? Several aspects could be realized:

- Standard-based formats of the encoding documents, metadata, and annotations were developed.
- A framework for unified access to linguistic services as well as dictionaries and other reference materials could be implemented.
- Tools were created to provide semantic access to sources, visualization of relations between data and metadata, and facilities for the production of new data.

As shown earlier, some of these tools are made available freely on the Internet. But what is even more is the fact that a large body of texts has been structured according the developed formats and made publicly available. Together with the tools integrated into the Archimedes working environment, the project contributed to the realization of a cooperative, research-driven open digital library. Yet the full realization of the original vision will require further work. The next project phase should be focussed on the following areas - besides, of course, an ongoing and substantial enlargement of the corpus of source texts which will form the body of knowledge for future research in history of mechanics and beyond:

4.1 Language technology

Research in natural language processing has resulted in a proliferation of tools for automatic morphological analysis of particular languages. However, there is no way to bring the data produced by these tools into a browsing or editing environment in the absence of standard formats and protocols. What is needed is the development of linguistic middleware enabling user agents to interact with heterogeneous sources of linguistic data

4.2 Semantic linking

Scholarly research requires the creation and utilization of meaningful links between source materials. Ordinary HTML links are unidirectional and cannot differentiate between types of semantic relations. Digital library applications must be developed that achieve the ideal of a "semantic web" by exploiting the potentials of relevant standards such as XML Linking Language.

4.3 Content creation

Central to the notion of a digital research library is human analysis and annotation of source material. The current paradigm of the Web distinguishes the creation and browsing of content as fundamentally separate activities. In a next-generation framework the browsing and creation of content must be more closely integrated.

4.4 Distributed resources

The free exchange of ideas is a crucial feature of research. Yet the current client-server architecture of the Web limits the free flow of information. Digital libraries must mover toward a more fluid distributed or peer-to-peer network model.

References

[1] The Archimedes Project Server in Berlin
http://archimedes.mpiwg-berlin.mpg.de
[2] The Archimedes Project Server at Harvard
http://archimedes.fas.harvard.edu
[3] MPIWG Workbench.
[4] Harvard Workbench
http://archimedes.fas.harvard.edu/