Constructing the InterPARES Thesaurus: A Vocabulary Tool for Diverse Research Communities

Nadav Rouche
US-InterPARES, UCLA Department of Information Studies (http://is.gseis.ucla.edu/), GSE&IS Building, Box 951520, Los Angeles, CA 90095-1520

InterPARES Project Overview [1]

The International Research on Permanent Authentic Records in Electronic Systems (InterPARES) aims to develop and articulate the concepts, principles, criteria and methods that can ensure the creation and maintenance of accurate and reliable records and the long-term preservation of authentic records in the context of artistic, scientific and government activities that are conducted using experiential, interactive and dynamic computer technology. Scholars in the arts and sciences, archivists, artists, scientists, industry specialists and government representatives from around the world are working together to undertake the challenge presented by the manipulability and incompatibility of digital systems, technological obsolescence and media fragility and to guarantee that society's digitally recorded memory will be accessible to future generations.

Vocabulary Tools

When research is carried out by a multidisciplinary and multicultural team that spans fifteen fields of inquiry and twenty countries, the precision and consistency of the terminology used in the course of the project is vital to the success of the research. For example, several terms that are key to this research refer to different concepts in each disciplinary and cultural environment involved, while similar concepts are expressed by different terms. The Terminology Team of the InterPARES project is responsible for researching all terms proposed for official use by each research unit within the project, and accepting or rejecting them on the basis of clarity and consistency with the other adopted terms in the various disciplinary and cultural contexts.

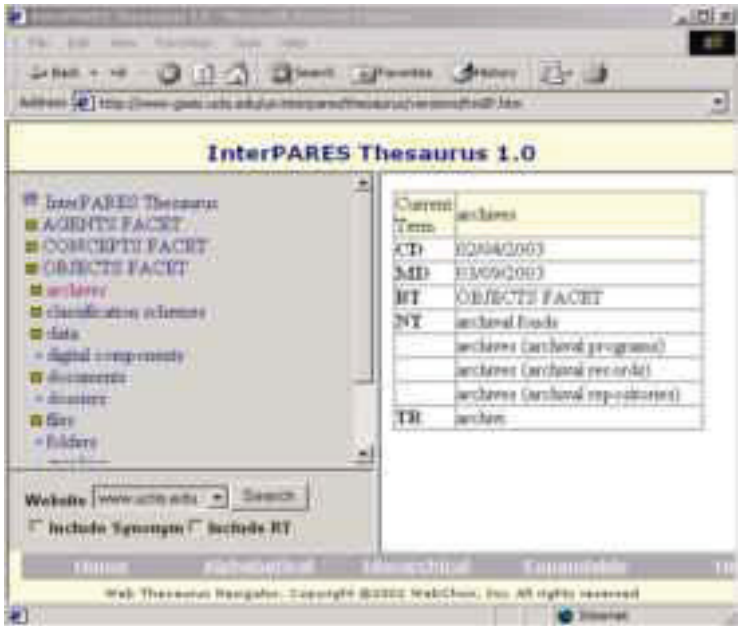
In order to reach its goal, the Terminology Team is responsible for developing a number of vocabulary tools: (i) a ***Register of Terms and Phrases*** to be selected from a corpus of InterPARES documents and the UBC Project Glossary. The register is a collection of working terms; (ii) a ***Glossary*** that will provide logical or conceptual definitions of the words and phrases in the register as they are to be used for working purposes within InterPARES, to provide for consistency. As such, the glossary establishes the preferred definition of a term within the scope of the project; (iii) consequently, developing an InterPARES ***Dictionary*** that will provide discipline-specific logical or conceptual definitions of the terms or phrases in the glossary (as well as additional terms and phrases that are not in the glossary) as they apply to the various disciplines. Thus, the dictionary includes both the preferred definition of a term as established in the glossary along with additional definitions of the term; (iv) and finally, developing an InterPARES ***Thesaurus*** which captures the relationships between most of the terms found in the other vocabulary tools and based upon the preferred definitions.

The InterPARES Thesaurus (beta version: www.gseis.ucla.edu/us-interpares/thesaurus/versions/home.htm)

The InterPARES ***Thesaurus*** is a web-based monolingual thesaurus built primarily using the ANSI/NISO Z39.19-1993 *Guidelines for the Construction, Format, and Management of Monolingual Thesauri.* The thesaurus development software selected for this project is TCS-8 from WebChoir (www.webchoir.com). The thesaurus interface supports multiple views of the thesaurus data (alphabetical, hierarchical, and

[1] <www.interpares.org>, and see also the US InterPARES Project <www.gseis.ucla.edu/us-interpares/index.htm>

expandable views) providing flexibility in browsing and looking up terms within the thesaurus by using complementary strategies. The thesaurus also supports querying external databases from a single interface, and semi-automatic query expansion.



Some of the research questions that were raised in the development process of the thesaurus were as follows:

- What are the strength and weaknesses of each vocabulary tool, and how do they complement one another? How can we reconcile different guidelines and standards governing the development of different vocabulary tools to provide maximal integration?
- How does developing thesauri in a digital environment impact the implementation of traditional thesauri building guidelines?
    - How do various types of interfaces support browsing, searching and learning terms within the thesaurus?
    - How, and to what extent, should syntactic rules of online search engines affect the formatting of the thesaurus in order to support proper query formulation?
    - How do various types of query expansions affect information retrieval performance?
- How can the thesaurus itself be used as a collaboration tool to support communication?
- What rules and procedures should be followed to develop, revise, and add terms to the thesaurus?
    - What should be the form and components of a thesaurus entry?
    - What should be the scope of the thesaurus and what languages should it support?
    - What are the criteria for determining hierarchical (broader/narrower term), associative (related term) and equivalence (use for) relationships between the terms?
    - How can researchers build a consensus among various research communities around the choice of terms and their placement in the hierarchy?

Project Funding: National Science Foundation and the National Historical Publications and Records Commission