Long-Term Electronic Storage, Retrieval and Use of Complex Engineering Information

T. MENZIES[a], V. MUCINO[b], J. MOONEY[a] AND T. MCGRAW[a]

[a]Lane Department of Electrical Engineering and Computer Science, West Virginia University, Morgantown, WV 26506-6109, USA

[b]Mechanical and Aerospace Engineering, West Virginia University, Morgantown, WV 26506-6109, USA

OUR QUESTIONS

Imagine the scene: some widely-used devices fails decades after it was designed. A new one must be designed.
Engineers turn to the archival storage to learn what they can from the old design. What problems would they encounter? How can we support them in that activity?

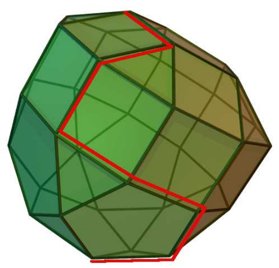
CHALLENGES

Based on surveys of engineering data sets, we assert that it is highly unlikely that a decades-old design exists in one particular format. Rather, technical documents like those seen in the STEP/EXPRESS format are stored along side a much larger set of supporting documents in multiple formats. A recent study concluded that :

- 80 percent of business is conducted on unstructured information.
- 85 percent of all data stored is held in an unstructured format.
- Unstructured data doubles every three months.

That is, if we can learn how to understand large heterogeneous collections that include STEP/EXPRESS knowledge as well as numerous other products in a wide variety of formats, it would be possible to reason and learn from a very wide range of data.

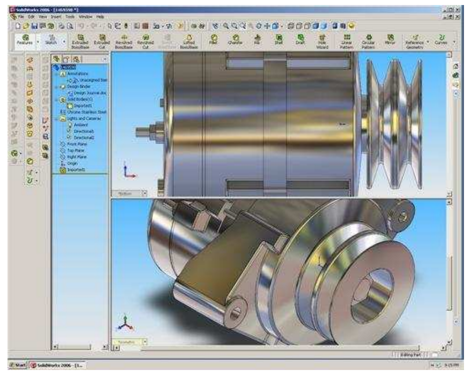
APPROACHES



- We need to understand *what information is lost*, using current methods, when engineers document their designs and those designs are added to a repository. This is the role of the *Tools Assessment* part of our work and is the current task of Victor Mucino and Jim Mooney.
- Our results to date show that we need to increase the amount of contextual and associational knowledge captured with STEP designs when we add those designs to an archive. Various *semantic visualization* tools will be useful for this work and this is the current task of Tim McGraw.
- Even with our best knowledge capture tools, it is highly likely that some percentage of the technical documents in the archive will be missing associational and contextual knowledge. Information retrevial methods can be used to recover part of that missing knowledge. This is the task of the *anything browser* work by Tim Menzies.
- Lastly, looking *over the horizon*, there are many other tools being developed to enhance capture of semantic information; e.g. the *semantic web* work. Exploring this next generation of tools is the task of Jim Mooney.

1 TOOL ASSESSMENT

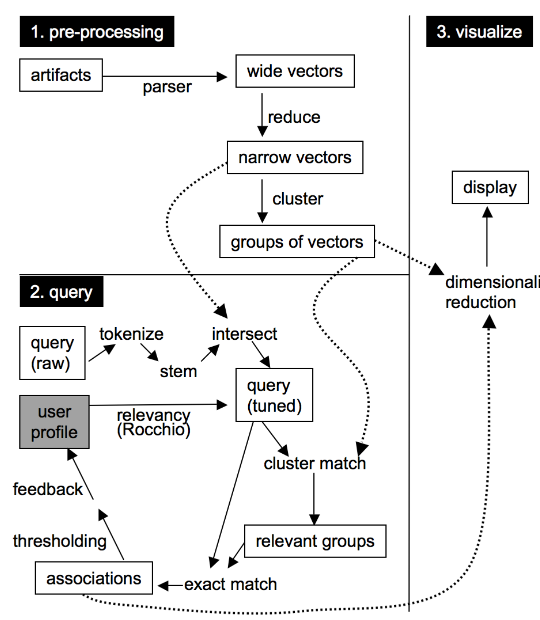
The STEP format (ISO-10303) for data exchange supports interoperability between CAD/CAM tools. Yet only a few of the application protocols (APs) developed for STEP are widely supported by mainstream CAD/CAM Software and industry. We ask: how much relevant engineering data is lost through the use of STEP for archival purposes?
To answer this question, we are building a framework architecture to provide a test-bed for using STEP APs in conjunction with various commercial CAD/CAM software which support STEP. Using the framework, we exercise various scenarios that require the generation and transmittal of data and information amongst engineering tasks. We then ask "how much of the information in those scenarios is captured in STEP?"



2 THE ANYTHING BROWSER

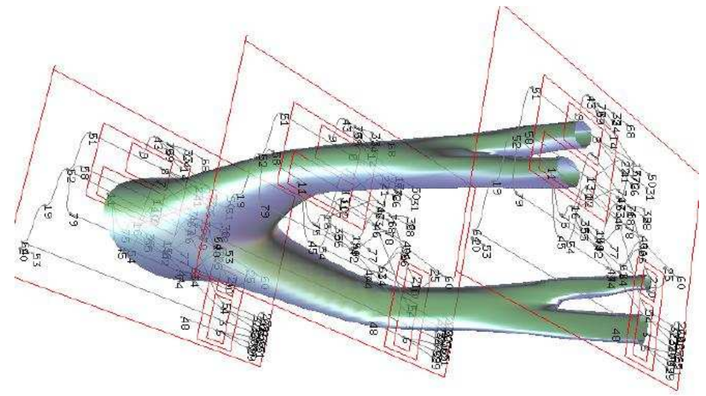
The anything browser is a *recommender system*. Automatic constrast set and classification tools assist the user in finding relevant materials and excluding irrelevant materials. Using the browser, some of the lost contextual and associational knowledge of a design can be recovered. The browser supports users in their decision-making while interacting with large information spaces. They recommend items of interest to users based on preferences they have expressed, either explicitly or implicitly. Recommendation technology represents a new paradigm of search: interesting items find the user instead of the user explicitly searching for them:

> *[We are..] leaving the era of search and entering one of discovery. What's the difference? Search is what you do when you're looking for something. Discovery is when something wonderful that you didn't know existed, or didn't know how to ask for, finds you.* – M. O'Brien

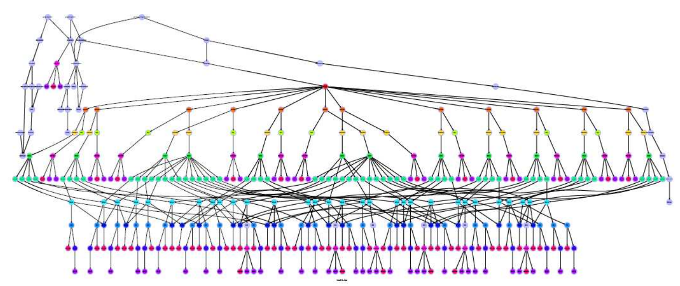


3 SEMANTIC VISUALIZATION

This work levers the anything browser work. Recall that the anything browser team will build parsers for a diverse range of materials. Using mostly automatic methods, that range of materials will be explored for clusters and common classifications. An alternative to the automatic filtering methods of the anything browser are visualization tools that display connections for users to browse. That is, in the anything browser, the filtering intelligence is in the algorithms. In the semantic visualizer, on the other hand, the filter intelligence resides with the operator while the tool offers alternative views.



The structure of large networks is typically conveyed by explicitly rendering nodes and arcs as geometric entities. We propose to additionally utilize color and texture to convey structural information.



The graphical browser will permit visualization of the hierarchical and semantic relations found in large document databases. New visual representations of graphs will be developed to enhance understanding of complex datasets. Fast response to user input will be achieved by exploiting the parallel processing power of the modern graphics processing unit (GPU).

4 OVER THE HORIZON

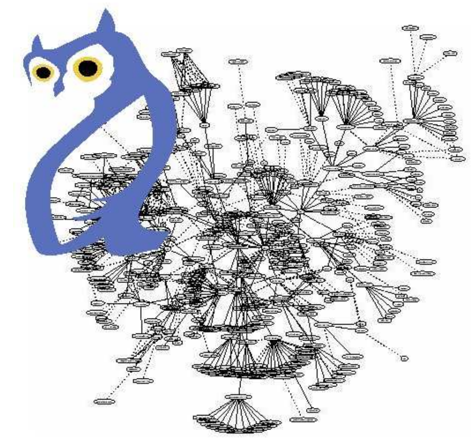
This project will continue to track and analyze STEP as it is evolving, and to identify future development most relevant for long-term data retention (LTDR). The team will develop proposals for extensions for STEP that may be presented to STEP developers for consideration.
One area of active investigation among STEP researchers today is the use of alternate representation schemes that may capture additional semantic information, such as the OWL Web Ontology Language. Participation in this study will determine the suitability of OWL or related forms for capturing engineering information.
It is also recognize that STEP is not the only representation scheme for engineering LTDR. An evident weakness in the use of STEP for archiving engineering product data is that, with a few notable exceptions, STEP is not being widely used as a representation for current projects. Among the reasons for this are:

- most aspects of STEP are not widely known,
- the actual specifications are not available at reasonable cost, and
- there are few tools to enable the use of the full range of STEP product data.

5 SEMANTIC WEB

Much of the proposed research is focused on the STEP specifications as represented in the EXPRESS specification language. However, it is expected that the work performed in 2008 will result in a detailed analysis of the STEP specifications as they exist today, and their suitability for long term retention of engineering data. This project will continue to track and analyze STEP as it is evolving, and to identify future development most relevant for long-term data retention (LTDR). The team will develop proposals for extensions for STEP that may be presented to STEP developers for consideration. One area of active investigation among STEP researchers today is the use of alternate representation schemes that may capture additional semantic information, such as the OWL Web Ontology Language1. Participation in this study will determine the suitability of OWL or related forms for capturing engineering information. It is also recognize that STEP is not the only representation scheme for engineering LTDR. An evident weakness in the use of STEP for archiving engineering product data is that, with a few notable exceptions, STEP is not being widely used as a representation for current projects. Among the reasons for this are:

- most aspects of STEP are not widely known,
- the actual specifications are not available at reasonable cost, and
- there are few tools to enable the use of the full range of STEP product data.



FOR MORE INFORMATION

Contacts:

1. Tool Assessment: Victor Mucino, Ph.D. (victor.mucino@mail.wvu.edu)
2. Anything Browser: Tim Menzies Ph.D. (tim@menzies.us
3. Advanced Graphics Tim McGraw, Ph.D. (Tim.McGraw@mail.wvu.edu)
4. Over the horizon: James Mooney, Ph.D. (James.Mooney@mail.wvu.edu)

SUPPORT

This research as carried out at West Virginia University under a contract with the National Energy Technology Laboratory with funding from the National Archives Electronic Records and Archives.