IMREC: A REFERENCE COLLECTION FOR INFORMATION MANAGEMENT AND RETRIEVAL IN ENGINEERING (IMRE)

Tomo CEROVSEK, Asst. Professor and Department Vice Head, tomo.cerovsek@fgg.uni-lj.si
University of Ljubljana, Faculty of Civil and Geodetic Engineering, Slovenia

ABSTRACT

An open-access edited repository providing a long-term scientific reference collection for research and development (R&D) in Information Management and Retrieval in Engineering (IMRE) is presented. IMRE goes beyond textual information retrieval (IR), as it addresses vast volumes of project-specific, task-centred, non-/semi-/well-structured engineering alphanumeric, graphical, 3D geometry and model data (e.g., analytical models, time-history series, analysis results and BIM). To facilitate the R&D in IMRE and to avoid redundant preparation of test data for large-scale test-beds IMRE Reference Collection (IMREC) is proposed. IMREC will provide a set of relevant queries, reference collections and procedures that will make IR benchmarks repeatable, comparable and interoperable. The purpose of this paper is to present the approach and to invite peers to make use of and to contribute to IMREC.

IMREC uses an R3-M6 approach, where R3 stands for Relevant, Reachable and Representative and M6 stands for Measurable, Multi-standpoint, Multi-application, Multi-phase, Multi-level-of-detail and Multi-lingual. Initially, we reviewed more than 70 existing test collections, developed a Dublin-Core repository and primed sample data. IMREC will serve for the R&D in IMRE to better address engineering information needs, improve communication in R&D and allow for technology transfer.

Keywords: information management, information retrieval, reference collections, aec workflows, bim

1 INTRODUCTION

The success or failure of an engineering business process depends on the efficiency and effectiveness of the creation, exchange and use of information throughout the project lifecycle. Because the amount of engineering data is growing (because of much easier authoring of digital content and an increasing amount of data produced by machines, e.g., sensory devices), the need for better information management and retrieval in engineering (IMRE) is growing and this perpetuates related R&D.

IMRE also addresses the retrieval of 3D, BIM and engineering models, which require new approaches to indexing, crawling and querying (Figure 1). Development and testing would be facilitated if representative engineering data sets, e.g., IFC models or time-history data, were freely available to researchers and developers. Currently, time and money are spent on the redundant preparation of collections, which are later not available to others to validate and verify research results.

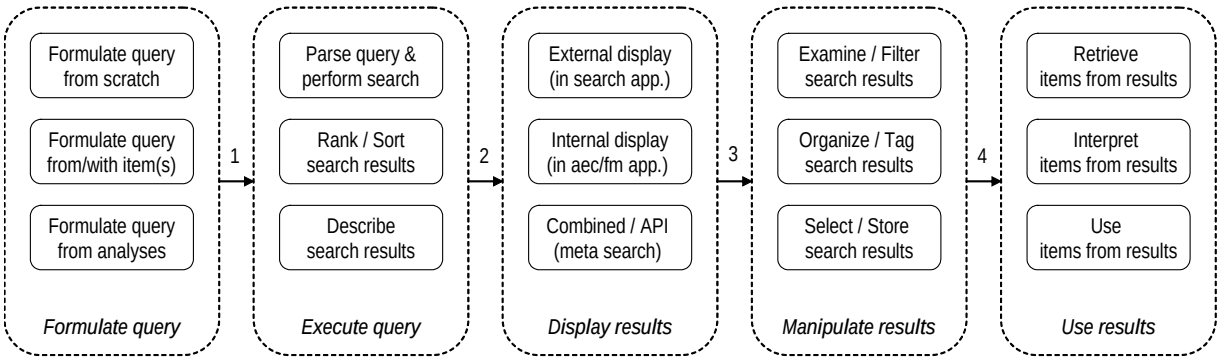

Figure 1: Querying of information on 3D geometry, BIM and other project specific information (Cerovsek 2008) is of vital importance for IMRE and could be facilitated with reference collections.

1.1 Statement of the problem

One of the main barriers that had slow down the development on querying techniques initiated by the author, see Figure 1 and Table 1, was the lack of relevant and representative data (due to volume, access rights or IPR). To illustrate the type and volume of information relevant to the IMRE, we list some important facts. The total volume of information that must be "consumed" in all project phases is staggering. The amount of mandatory information alone, i.e., building codes and standards, exceeds thousands of pages, which cannot be processed and used without information management tools. The volume of project-specific information is also barely manageable. Some relevant statistics follow:

- **4.000 (1.5 GB) documents per project.** A case study (Caldas 2002) showed that 16 project team members generated around 4 000 document files (1.5 GB) of project data.
- **800.000 (300 GB) documents per year.** A large manufacturing/construction company that makes over 200 projects per year produces more than 800,000 documents (300 GB) per year.
- **1 drawing per 10 sq meters or 2800 drawings for a bigger building per project.** This set of drawings also contains versions that must be managed and distributed.
- **1 BIM model may contain millions of polygons (from 10 to 200 MB/model).** Building models may contain thousands of elements, which may be represented by millions of polygons
- **GIS data (100s of GB).** A 60% annual growth rate of GIS data is estimated for the future.

Additionally, empirical data from experiments and devices in smart buildings are generated as follows:

- **16x6.700 double-float numbers per simple experiment (GBs).** An average experiment with inductive meters produces 1.000 to 10.000 double-float numbers per second per channel. The total amount of data depends on the duration, frequency and number of measurement sites.
- **1 double-float/minute across several years.** The amount of measurement data depends on the data streams, how long they operate and frequencies of measurements per time period. The use of sensory devices in every-day buildings is additional contributing vast volumes.

Problem: *Despite the fact that the volume of engineering data is increasing, researchers do not have sufficiently large data sets to permit scientific evaluation and comparison of IMRE research results.*

1.2 Premises and the objective of the paper

The premises that are addressed by the objectives of this paper are as follows:

- The information that must be handled in AEC sector is already voluminous and complex.
- The information generated by software and hardware is not sufficiently/efficiently managed.
- The amount of information that is relevant for decision making in the life-cycle is growing.
- The information is sensitive and wrong interpretations can lead to extra costs, delays and risks.
- The engineering information-seeking behaviour and management requires new tools.

The objective of this paper is:

- To give an overview of existing reference collections and their role in the development of IR.
- To provide a description of content types and the character for the reference collection.
- To present the initial organization of the content and procederes for the reference collection.
- To invite peers to make use of reference collection and contribute to the reference collection.
- To advance the development in information management and retrieval in engineering.

Solution: *A reference (test) collection, which is maintained for the purpose of the study of engineering information management and retrieval of AEC-specific information (including BIM/IFC/IFD models), in all project phases would significantly improve IMRE research, evaluation and technology transfer.*

1.3 AEC vertical search

This Section provides an overview of relevant content types and formats for AEC vertical search, which is very important for IMRE. Search engines may be divided by the type of content that they have access to and are developed to index and search the content both vertically and horizontally. We commonly use broad, general-purpose search engines, although some searches focus on data in a specific domain. This relatively recent trend in the search community is called a *vertical search*.

In general, search facilities are usually separated into four distinctive, domain-independent categories that provide *horizontal coverage* of the search space and applications as follows:

- Web search (based on an index of publicly available data)
- Enterprise search (based on the collective index data from an organization, e.g., Google SA)
- Desktop search (based on an index of personal, individual desktop data)
- Database search (based on the data that are stored in relational or object-oriented databases)

Above types of search engines index different formats and make use of different querying techniques. Whereas general-purpose search engines use rather simple, natural language-querying techniques, vertical search engines require new techniques. The table below presents an overview of AEC-specific content types and information retrieval issues related to the querying techniques and query carriers that should be addressed for IMRE.

Table 1: A review of content types and querying techniques for AEC (adapted from Cerovsek 2008)

	AEC content type				
	Text *word processing* codes, specifications, contracts, correspondence	**Structured data** *spreadsheets/db* quantity survey data, point cloud, time history	**Still imagery** *vector or raster* drawings, photos, diagrams, graphs, sketches	**Audio and Video** *audio or video* recordings of real or digital world, time lapse	**Engineering** *I/O engineering* analysis, simulation, process and product models
Query by ...	**Type of query to be used to find AEC content type**				
... Syntax	X	X			
... Template	X	X	X	X	
... Example	X	X	X	X	X
... Co-occurrence	X	X	X	X	X
Query carrier	**Type of query carrier that can be used to find query**				
Text	X	X	X	X	X
Structured	X	X	X	X	X
Still imagery	X		X	X	
Video	X		X	X	X
Sound	X			X	X
Model	X	X	X	X	X
Combined	X	X	X	X	X

For IMRE, it is important that we understand and consider the whole lifecycle of information . Therefore, we are particularly interested in the role of IMRE in workflows that are crucial for the success or failure of engineering business processes. Although searching can be a stand-alone task, it is most frequently a part of a business process that includes several steps. Note that workflows usually include several content types, which cannot be handled with existing reference collections.

2 REVIEW

In this section, we provide a brief review of some of the most relevant existing reference collections with basic background information. Initially, we reviewed over 70 reference collections and organizations and their practices. The analysis of reference collections is divided according to three ownership categories to which representative reference collections belong:

- **Standardization organization collections.** The most prominent example, co-sponsored by the National Institute of Standards and Technology (NIST), is the Text Retrieval Conference (TREC), founded in 1992 as part of the TIPSTER Text program.
- **Research community collections.** Research associations, such as IEEE or the Association for Computing Machinery (ACM) and its special interest groups (e.g., SIGKDD, Special Interest Group for Knowledge Discovery and Data Mining), latter is one of the leading professional organizations of data miners that also organizes an annual competition, the KDD Cup.
- **Proprietary collections.** Proprietary collections are owned by state agencies, individuals like Castillo (2006) or Koegh (2006), commercial companies (e.g., Google test collections) or other resources that have limited or no connections to the above-mentioned organizations.

For each category above, we describe the motivation for the development and maintenance of the reference collection, its organization and processes and the reference collection contents.

2.1 Motivation for developing reference collections

The incentives for the development of reference collections vary; however, organizations and individuals have, in general, similar agendas. From the point of view of organizations, reference collections are developed to initiate new developments and to test algorithms or applications related to information-retrieval issues (e.g., indexing, ranking, summation, summarization, clustering and evaluation of interfaces) and to compare different approaches, as well as to compete with and differentiate large and relevant collections. Both TREC and ACM special interest groups prepare reference collections to organize annual information-retrieval conferences and competitions, the purpose of which is to support and to conduct further research in the information-retrieval community. Although collections may provide diverse datasets, they are often too small to make claims about the efficiency of techniques. From the point of view of individual researchers, reference collections serve three main purposes: (1) for research, (2) to publish conference papers and peer-reviewed journals articles or patents and (3) as a reference and got referenced based on TREC.

Standardization organization collections. TREC goals are to provide a large-scale evaluation of text-retrieval methodologies; to increase communication among industry, academia and government; to speed the technology transfer; and to increase the availability of appropriate evaluation techniques. The effectiveness of retrieval systems approximately doubled in the first six years of TREC. According to a survey by Rowe et al. (2010), TREC was important specifically for research and publishing.

Research community collections. SIGKDD aims to provide the premier forum for the advancement and adoption of the "science" of knowledge discovery and data mining. Other communities stress that redundancy in operations related to the preparation of test collections is reduced and that more time can be dedicated to novel research, while the level of comparison is highly increased. Many other relevant communities, such as SIGIR and SIGMOD, use a similar approach.

Proprietary collections. The motivation for the development of proprietary collections is similar. For example, Castillo (2006) wanted to create a large, clean collection without classification errors, with uniform random sampling over a dataset, while covering broad topics with open access for researchers. Other developers (NeesGrid) aim to use and access the repository, to learn about experiments and their results and to discover new information in a repository populated with experimental data.

2.2 Organization and process

The importance of reference collections is illustrated in Figure 2, which shows how organization and the development and use of reference collections. In general, collections are most often used to organize advanced developments, competitions and workshops and to unify evaluation.

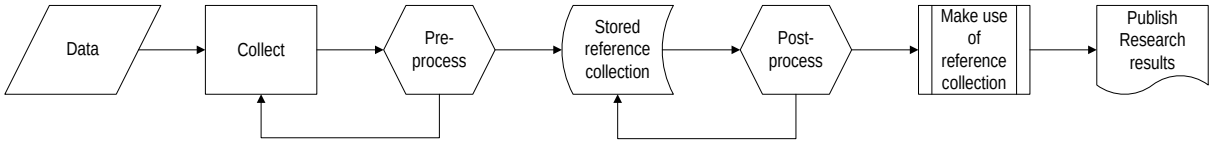

Figure 2: Typical development and use of a typical reference collection

Standardization organization collections. A program committee consisting of representatives from government, industry and academia oversees TREC and provides a test set of documents and questions for each TREC collection. Participants run their own retrieval systems on the data and return a list of the retrieved top-ranked documents to NIST. Next, NIST pools the individual results, judges the retrieved documents for correctness and evaluates the results. The TREC cycle ends with a workshop that is a forum for participants to share their experiences.

Research community collections. ACM provides and serves its members and the computing profession with leading-edge publications, conferences and resources. ACM includes 36 special interest groups (SIGs) that address different aspects of computing. SIGs organize competitions based on tracks that they provide. Once the competition is announced, the registration opens; once the data are available, along with tasks and rules, the competition ends on a predefined date, when winners are announced and results presented at the workshop.

Proprietary collections. In many cases, available test collections are insufficient and individual researchers and organizations must collect, build, label and manage their own test collections. Castillo et al. (2006) report on the creation of a Web-spam database with Web crawling, elaboration of Web spam guidelines and classification interface, labelling and post-processing.

2.3 Content of reference collections

The collections shall be sufficiently large to realistically reflect real-world situations. The TREC test collections and evaluation software are available to the retrieval research community to test their own search facilities. Content details are presented in Tables 2 and 3. Some relevant reference collections come from various domains, such as an open-access collection of earthquake engineering data with live and historical sensory and general public and scientific location-related content. An example that combines proprietary, standardization and community data can be found at http://earthquake.usgs.gov.

Standardization organization collections. TREC supports large-scale evaluations of the retrieval of English and non-English (Spanish and Chinese) multi-lingual and open-domain answers to questions and content-based retrieval of digital video (more details are shown in Table 2). Furthermore, NIST (2011) and SHREC 2011 (Shape Retrieval Contest based on Generic 3D Dataset).

Research community collections. The KDD organizes contests based on large, heterogeneous, multi-disciplinary reference collections, which often combine human perception with practical importance, behaviour prediction, algorithms for scoring and ranking (some details in Table 3).

Proprietary collections. Several proprietary collections have had a strong impact on R&D. For example, van Rijsbergen (1994) created the Glasgow collection with 25k docs and Lewis (2004) compiled the Reuters collection for text categorization. CMU (2011) collected 750.000 face images for face recognition, Koegh (2006) created the UCR collection for classification, clustering and indexing the time series, and large geometric models (GIT 2011) are dedicated to the creators of new geometric algorithms at the Stanford Scanning Laboratory, which includes several 10^8 XYZ RGB cloud-point models. A database of 3D models is available at http://ir4aec.caece.net and on the Princeton site at http://shape.cs.princeton.edu/benchmark/index.cgi (see Table 4).

Table 2: Example of a standardization organization of TREC tracks (source: Rowe et al. 2010)

Area of IR Research	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	Collection content
Use of TREC Resources																	
Personal documents																	Blog
																	Spam
Retrieval in a domain																	Legal
																	Genome
Answers, not docs																	Novelty
																	Q&A
Web-based searching, large sizes																	Enterprise
																	Terabyte
																	Web
																	VLC
Beyond simple text																	Video
																	Speech
																	OCR
Beyond English																	x→{x,y,z}
																	Chinese
																	Spanish
Human itl.																	Interactive
Streamed text																	Filtering
																	Routing
Static text																	Queries
																	Ad Hoc

Table 3: Example of the research community collection SIGKDD (http://www.sigkdd.org/kddcup/)

KDD CUP Centre																	
Consumer recommendations																	Log files
Pulmonary embolism detection from img																	Images
Internet user search query categorization																	Queries
Particle physics and protein homology																	Profiling
Network mining and usage log analysis																	Log files
BioMed document and gene classification																	Genome
Molecular bioactivity and protein predic-																	Record sets
Online retailer website clickstream																	Log files
Computer network intrusion detection																	Log files
Direct marketing for profit optimization																	Queries
Direct marketing for lift curve optimiza-																	Profiling

Table 4: Example of proprietary collections from different relevant fields

Resource type	Description	Metadata	Volume
Text	Glasgow collection	Bibliographic	24,966 docs
	Reuters test collection	Bibliographic	21,578 docs
Structured data	UCR Time Series Page	Subject description	22 cat. 1188 sets
	Strong motion earthquakes	Triggered events/Region/Peaks	129 recordings
Still imagery	Multi-PIE—face database	Subject, illumination, view	750,000 images
	Out-text texture collection	Illumination, material	320 cat. 27054 images
Audio and video	Open video project	Metadata for each segment	3902 segments
Engineering data	Large geometric model	Faces/vertices	12 with 108 points

3 IMREC

The goal of IMREC is to facilitate R&D and to provide a base for better communication of research among stakeholders. This means that information management should include cross-disciplinary information workflows. The developments were initiated via an IMRE LinkedIn group that was established in 2009, and in January, 2011, members were invited to collaborate on IMREC.

3.1 Motivation for IMREC

The motivation for IMREC is, in many ways, no different from the incentive of existing reference collections, while the context is noticeably different. IMREC goals were:

- To assure repeatability, comparability, availability and interoperability of IMRE tasks.
- To remove barriers to research by making the content and procedures freely available.
- To enable easier discovery, indexing, searching, and reuse of process and product data.

3.2 Organization and process R3-M6

The organization of the IMREC is an open access reference collection to facilitate the process of collecting, building and sharing . The organization of engineering data may be classified by discipline, project phase, data structures, designed artefact and function to enable easier management and use of procedures. All collected data and performed benchmarks should provide a description of inputs, procedures, controls and outputs for each activity in the information flow, as described in Figure 3.

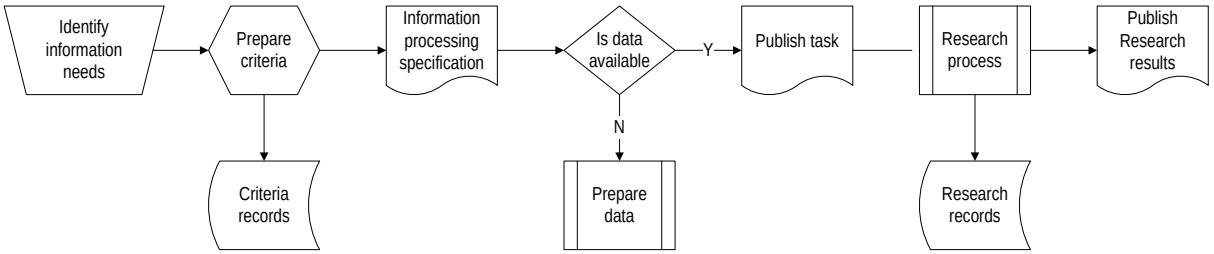

Figure 3: Typical preparation of the reference collection

The content available in the reference collection should meet three requirements:

- *R1 – Relevant*: Content and procedures should be relevant to engineering workflows.
- *R2 – Reachable*: The reference collection should be reachable to make research repeatable.
- *R3 – Representative*: Quantitative and qualitative criteria should be fodder for inclusion.

The results should be:

- *M1 – Measurable*: The information management and retrieval task should be measurable.
- *M2 – Multi-standpoint:* The information may be observed from different standpoints.
- *M3 – Multi-application*: The task may include a workflow from many applications.
- *M4– Multi-phase:* The information management or retrieval task may include more phases.
- *M5– Multi-level-of-detail:* The same information may be handled at different levels of detail.
- *M6– Multi-lingual:* Information management may involve different languages.

Controlled vocabularies (e.g., synonym rings, taxonomy, classification, thesauri and ontologies) are important to IMRE for consistency, communication structuring, labelling and retrieval. A major goal of vocabulary control is to ensure that each distinct concept refers to a unique linguistic form.

3.3 Content of the reference collection

There are three main relevant types of engineering information:

- *Project information* (data produced in all project phases: pre-, during and post-construction)
- *Regulatory information* (data relevant for compliance with regulations and standards)
- *Empirical information* (data (streams) from experiments and embedded sensors)

The topics covered. The collection of content includes theoretical knowledge of engineering design methodologies, empirical studies, engineering building codes and standards, project work (modelling, analysis and simulations) and natural disasters. Relevancy should be observed through the essential requirements of mechanical resistance and stability, safety in case of fire, hygiene, health and the environment, safety in use, protection against noise, energy economy and heat retention.

Descriptions of the collection. Metadata are organized according to the chosen Dublin Core data:

- Resource type: collection type, text, structured dataset, sound, video and engineering.
- Quantity and definition: description of items with indication of the strength(s) of the collection
- Format: digital file format, such as xml, txt, doc and a list of preferred digital file formats
- Ownership and access rights: An entity that has legal possession of the collection's access rights may include information regarding access or restrictions based on privacy or security.
- Item metadata: Author, title, publication year, summary or abstract, type of material, keywords or subject terms, classified (UDC), language, document format, number of pages/volume, publisher, copyright statement and document size/volume

Table 5: IMREC initial collection (http://imrec.deskriptor.si)

Resource type	Description	Metadata	Volume	Format
Text	Textual documents	Bibliographic	1939	PDF/XML
Structured datasets	Experiment recordings	Experiment information	139	ASCII
	Point-cloud data	Location/date and time	10^{12}	3DP
	Weather data	Location/T,H,V,R	10^{5}	ASCII
Still imagery	Earthquake images	Location/Structure/Failure	580	JPG
	CAD files	Embedded CAD metadata	650	DXF
	Renders	JPG header metadata	590	JPG
Audio and video	Fly-through videos	Fly-through videos/VR	280	QTVR
Engineering data	Time-history data	Analysis, experiments, device data	10^{4}	ASCII
	Building information models	Embedded metadata	480	IFC

The example bellow shows a possible use of IMREC data and procedure from Figure 2.

Example: *Point cloud TLS scan to BIM conversion competition*

- *Identification of information need*: Semantic conversion of a scan:
  - Convert point cloud of buildings data to CityGML/IFC
- *Criteria*: Ranking of results will be based on the
  - 40% geometry matching: the volume of the difference between CP and BIM models
  - 40% semantic mapping: the number of correct elements assigned
  - 20% time of creation: CPU time required for processing

We would continue with an *information-processing specification* that would include a detailed description of processing and, based on this description, identification of need and criteria, we would be able to decide whether an existing reference collection contains relevant data.

4 DISCUSSION AND CONCLUSIONS

The development of reference collections such as IMREC is a response to an increasing need for better information management and retrieval in engineering (IMRE). The scope of existing freely available reference collections is insufficient for AEC-specific content types and information management and retrieval; main problems are project specific context, volume and IPR issues. The lack of relevant and representative reference collections is one of the main barriers to advancement in research on IMRE. The researchers spend time and resources for redundant preparation of input, while their research results cannot be validated and verified because the same test data is not available to others.

IMREC reference (test) collections could provide a valuable source for research in IMRE. IMREC could be used by individual researchers, for group experiments, competitions and/or in combination with engineering workflows that could make use of several types of content and tools.

IMREC suggests the R3-M6 approach: Relevant, Reachable and Representative–Measurable, Multi-standpoint, Multi-application, Multi-phase, Multi-level-of-detail and Multi-lingual. IMREC will assure input for research and development with a potentially heavy impact on the development and advancement of intelligent engineering information-management tools, on the one hand, as it will allow end-users to express their information needs, on the other hand. Furthermore, IMREC will have the potential to bring together researchers and software developers and will enable the transfer of knowledge from academia to the industry. In the future, the work will focus on the following:

- Extension of the collaboration for the advancement and development of reference collections
- Promotion of the use of reference collections and attracting research communities, e.g., CIB
- Development of sustainable procedures and practices that will enable technology transfer

ACKNOWLEDGEMENTS

The contribution of the content by individuals and organizations is gratefully acknowledged. Special thanks are also owed to the individuals that supported initial activities, Thomas Liebich from AEC3 for assisting in the supply of initial sample building information models, many faculty members and students; and organizations like Riegel and Trimble that made possible the use of point cloud data.

REFERENCES

Arijit, S. & andrew, D. (1997). “Query By Templates: A Generalized Approach for Visual Query Formulation for Text Dominated Databases.” *Fourth International Forum on Research and Technology Advances in Digital Libraries (ADL '97)*. Los Alamitos, CA, USA, IEEE.

Rowe, B.R., wood, D.W., link, A.N., and simoni, D.A. (2010). “Economic Impact Assessment of NIST’s Text REtrieval Conference (TREC) Program” NIST, RTI Final Report NIST.

Borrmann, A., van treeck, C., rank, E. (2006). “Towards a 3D Spatial Query Language for Building Information Models” http://itc.scix.net/cgi-bin/works/Show?c71e

Caldas, C. H., soibelman, L and han, J. Automated Classification of Construction Project Docs J. Comp. in Civ. Eng. 16, 234 (2002), DOI:10.1061/(ASCE)0887-3801(2002)16:4(234)

_____, T (2008). On AEC Query Formulation techniques. ECPPM 2008, Taylor & Francis.

CMU (2011). The CMU Mutli-PIE Face Database. http://www.multipie.org/

Keogh, E., Xi, X., Wei, L. & Ratanamahatana, C. A. (2006). “The UCR Time Series Classification/Clustering Homepage” www.cs.ucr.edu/~eamonn/time_series_data/

Koh, E. Kerne, A. & Berry, S. 2009. Test collection management and labeling system. In Proceedings of the 9th ACM symposium on Document engineering (DocEng '09). ACM, New York, NY, USA, 39-42. http://doi.acm.org/10.1145/1600193.1600203

Lewis, D.D. (2004). “Test Collections” http://www.daviddlewis.com/resources/testcollections/

Nist (2011). “SHREC 2011”. http://www.itl.nist.gov/iad/vug/sharp/contest/2011/Generic3D/

Van Rijsbergen, K. (1994). Test Collections. http://www.dcs.gla.ac.uk/~keith/