Using KOS as a basis for text analytics and trend forecasting

NKOSS **Networked Knowledge Organization Systems and Services**
The 9th European NKOS Workshop at the 14th ECDL Conference, Glasgow, Scotland
10th September 2010

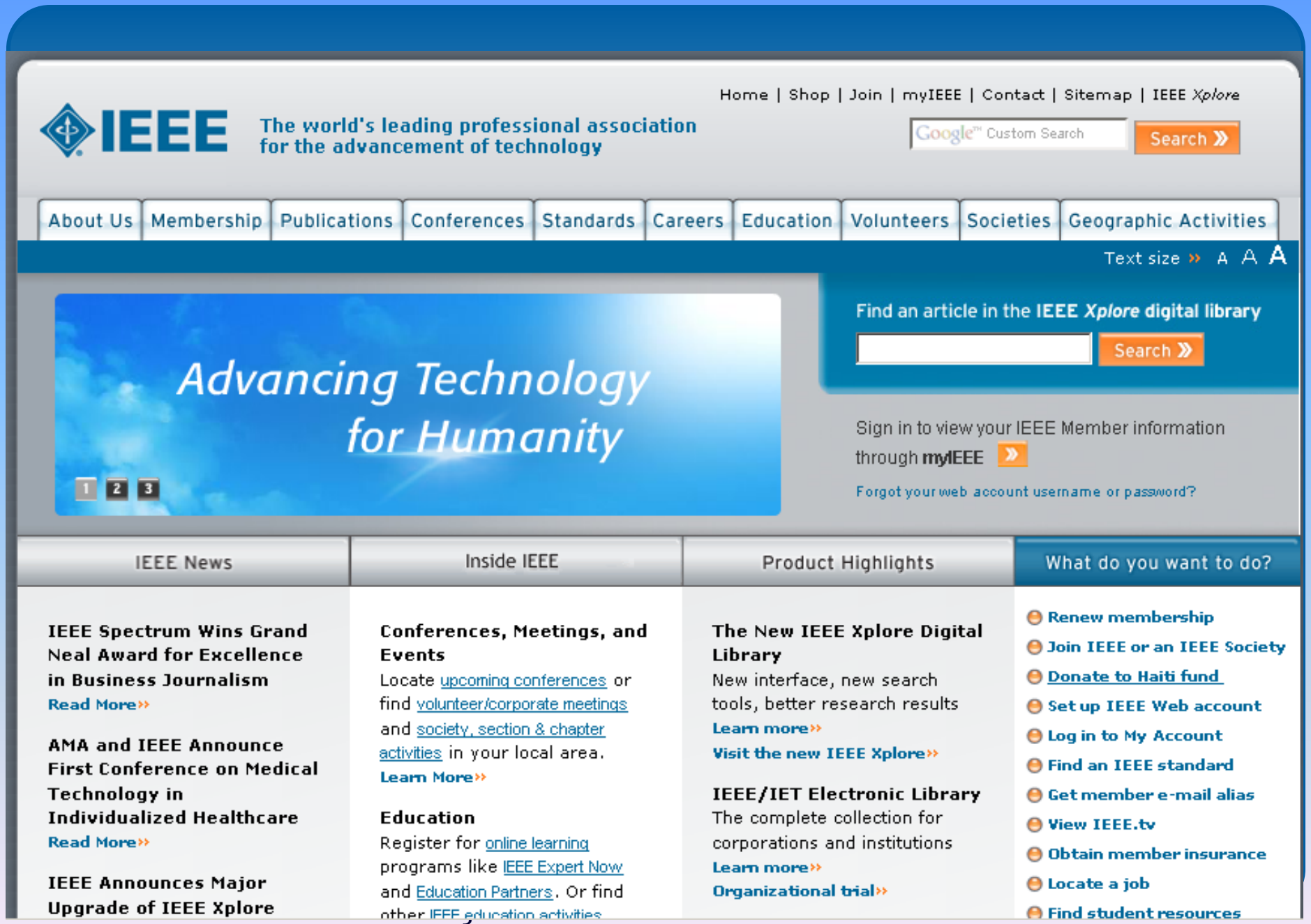
Marjorie M.K. Hlava
President
mhlava@accessinn.com

www.accessinn.com

Access Innovations.

Agenda

- IEEE Challenge
  - Where are our publication strengths?
  - What are the emerging topics?
  - Use our own data to address these question
- Access Innovation's Response
  - Expanding and Mapping the IEEE Thesaurus
  - Use term instead of text analytics to investigate
- Findings

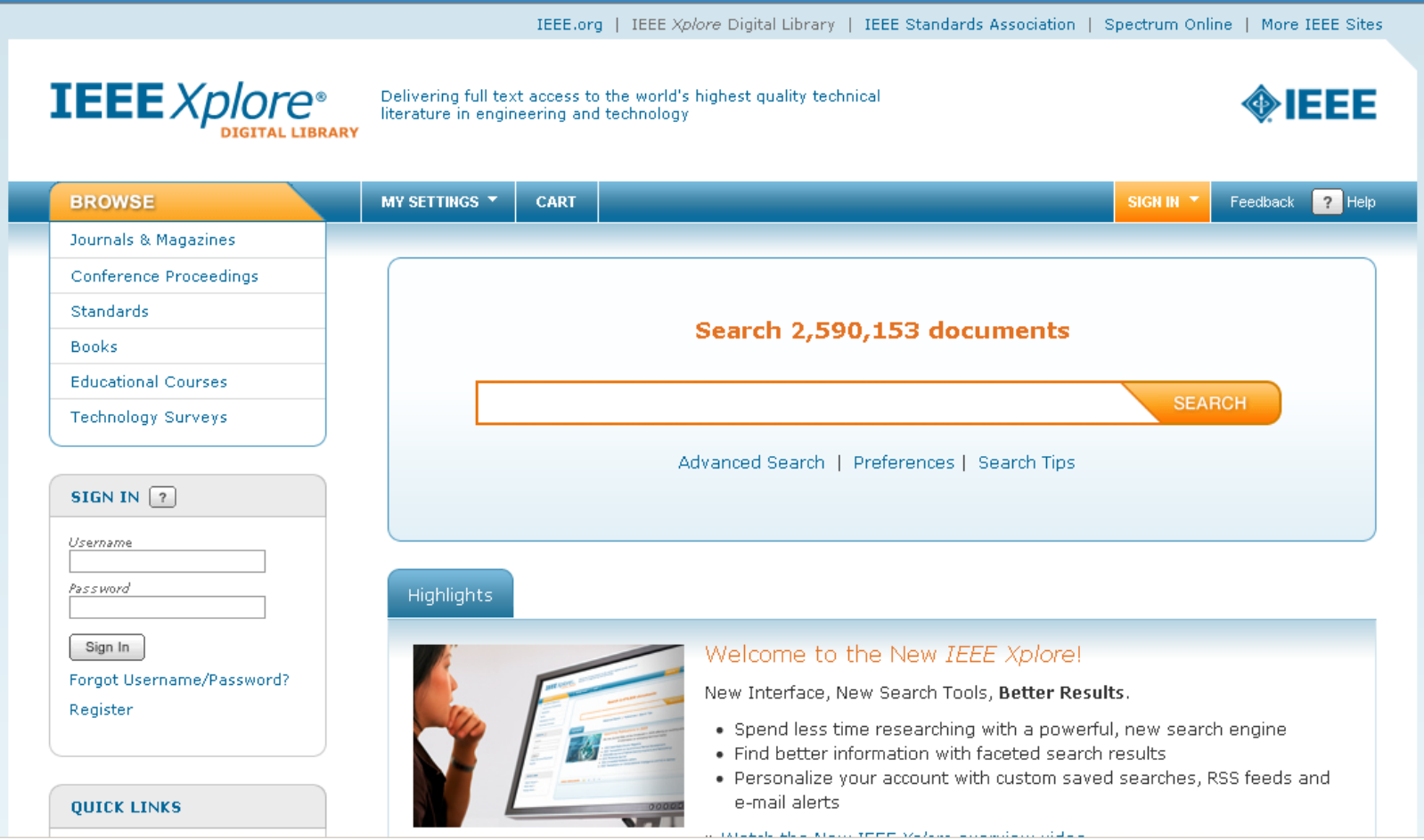


Access Innovations.

About IEEE…

- Founded in 1884, IEEE is the world's largest professional association advancing technology for the benefit of humanity.
- We publish 148 technical journals, transactions and magazines, sponsor nearly 800 conferences annually, develop technology standards, and support the professional interests of more than 400,000 members in over 160 countries.
- Members participate in 38 societies and 7 councils
- The **IEEE *Xplore*® digital library** provides access to IEEE journals, transactions, letters, magazines and conference proceedings, IET and other 3rd Party journals and conference proceedings, IEEE Standards and IEEE educational courses.
  - Over 2.5 million documents

The *New* IEEE Xplore



Access Innovations.

Specific Challenges

- Is there a way, using IEEE information, to forecast future direction?
- Where is the industry headed? What about by technology sector?
- Does our coverage match the IEEE mission and vision?
- Can IEEE become smarter about their data and potential markets using their collection in new ways?
  Are the societies publishing and talking about what their charter indicates they cover?
- What are the trends – are topics emerging/cooling?
- Can IEEE use technology and their own data to explore these questions while enhancing their data?

Access Innovation's Response

- SciTech Strategies' Maps
- Access Innovation's Tools
- IEEE Xplore data
- Test with several thesauri

Access Innovations.

Access Innovations / Data Harmony

- Founded in 1978
- Suite of Semantic Enrichment tools
- Updated the IEEE Thesaurus in 2005
- Built a rule base to auto index IEEE content
  - "90 % accuracy out of the box on journal data"*
  - "80% out of the box on proceedings data"*
- Auto indexed 1.2 million Xplore records
  - With the IEEE thesaurus terms rule base
  - With the MeSH rule base
  - With DTIC rule base

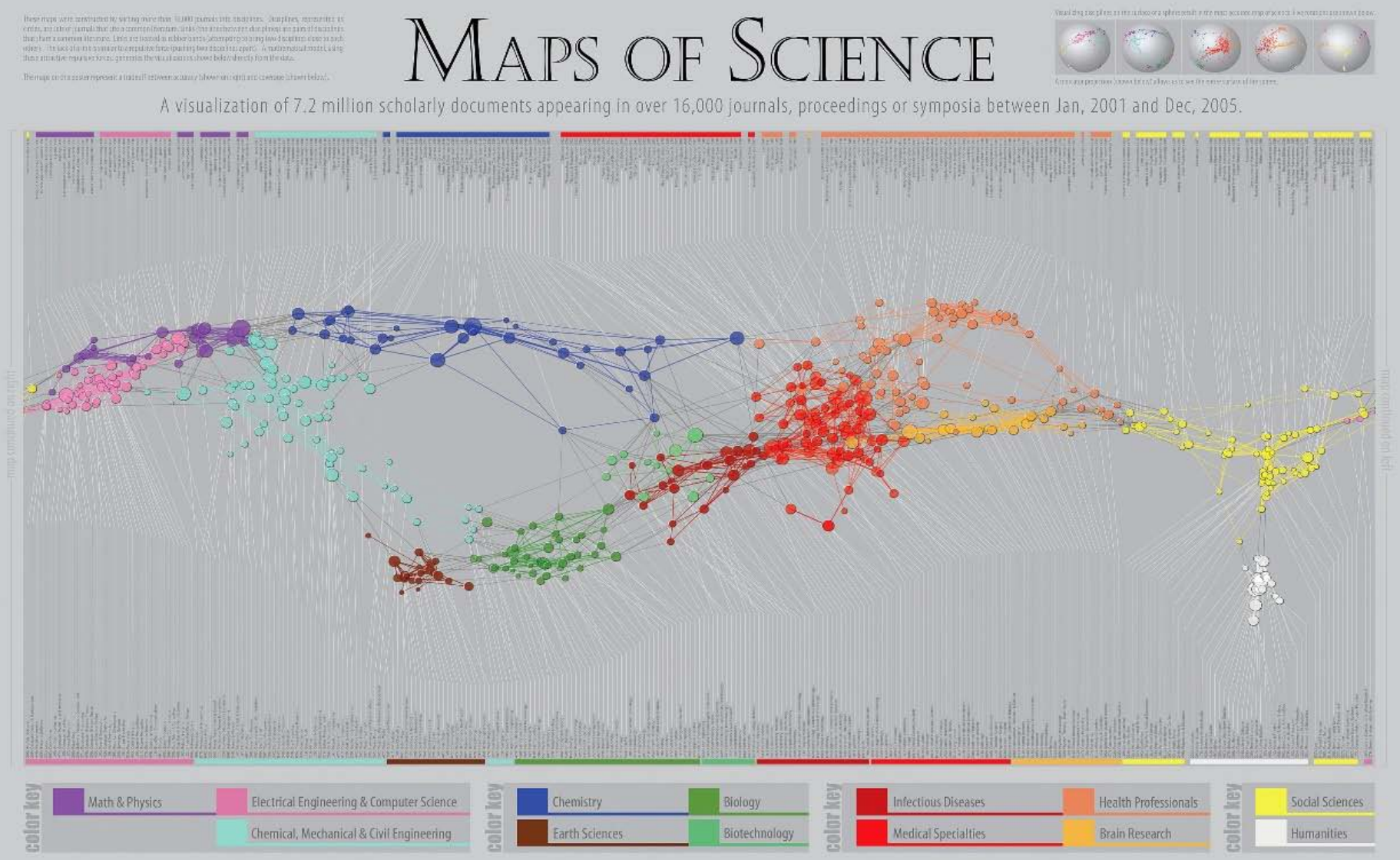
*Adam D. Philippidis, Manager, Indexing & Database Production, IEEE

Access Innovations.

SciTech Strategies, Inc.

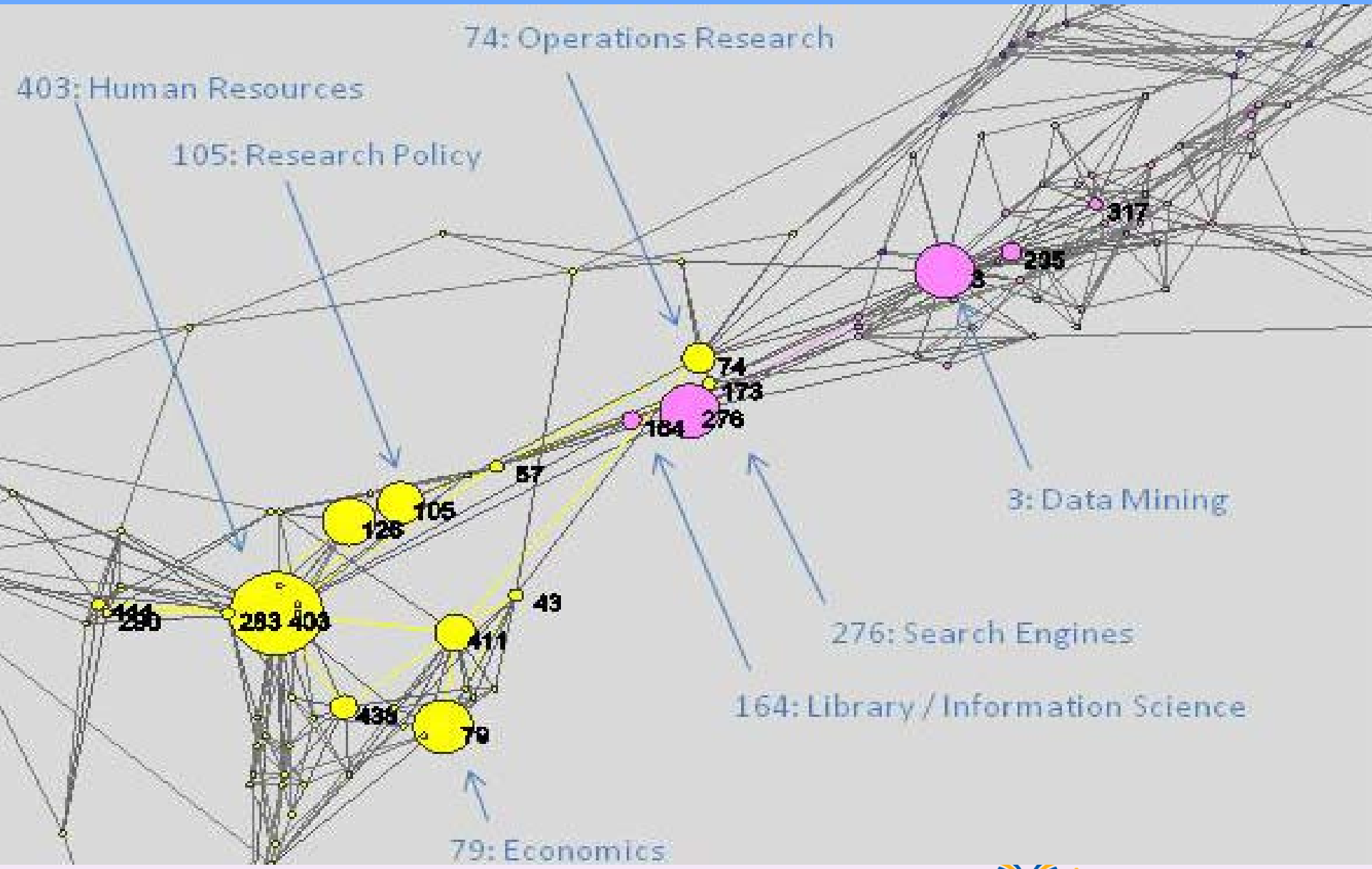
- Founded in 1982 (Center for Research Planning)
- Bibliometric Modeling of very large datasets
  - Thomson/ISI data (1982-2004)
  - Elsevier/Scopus data (2004-present)
- Focus on Accuracy
  - Disciplines > Research Communities > Researchers

Disciplinary Map of Science



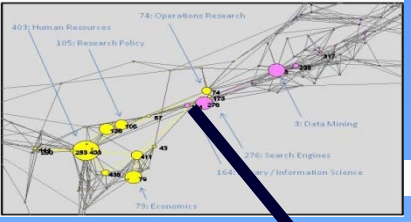
Access Innovations.

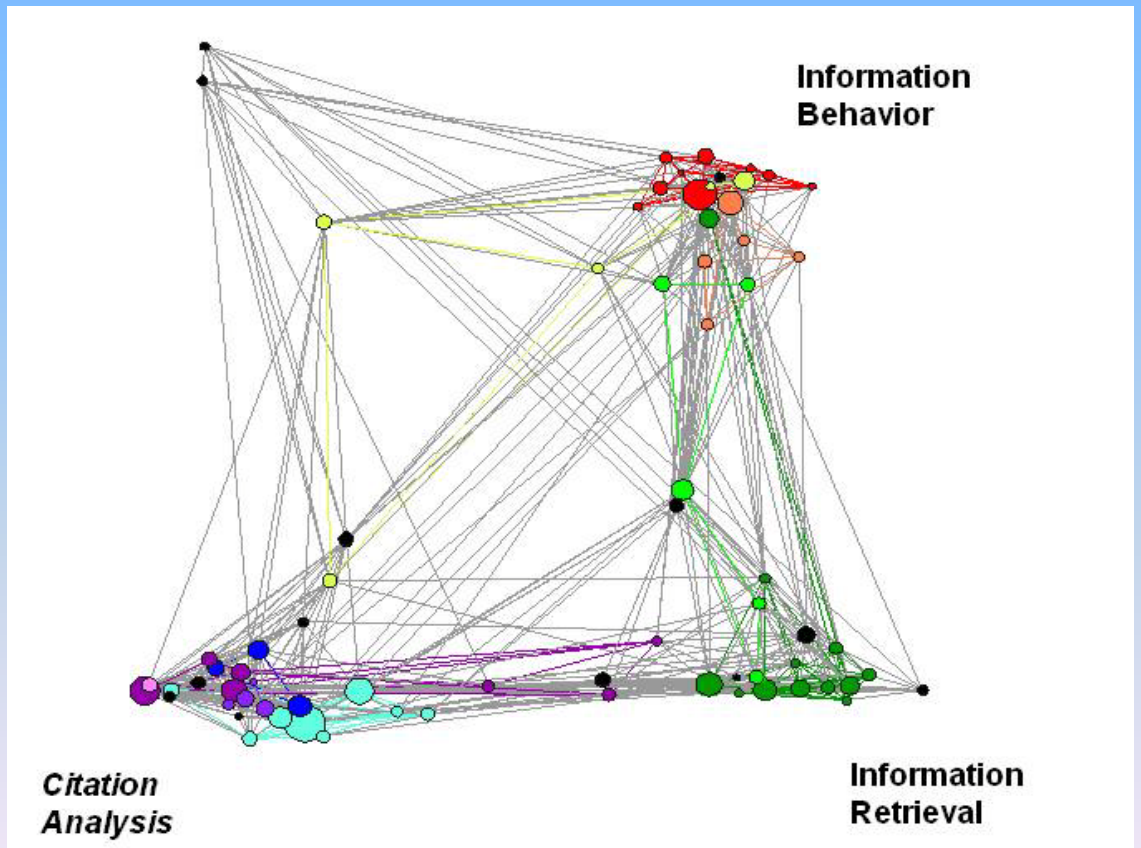
Relevant Disciplines for Science Mapping



Access Innovations.

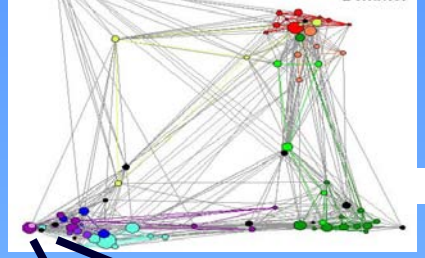
Research Communities in the Information Science Discipline

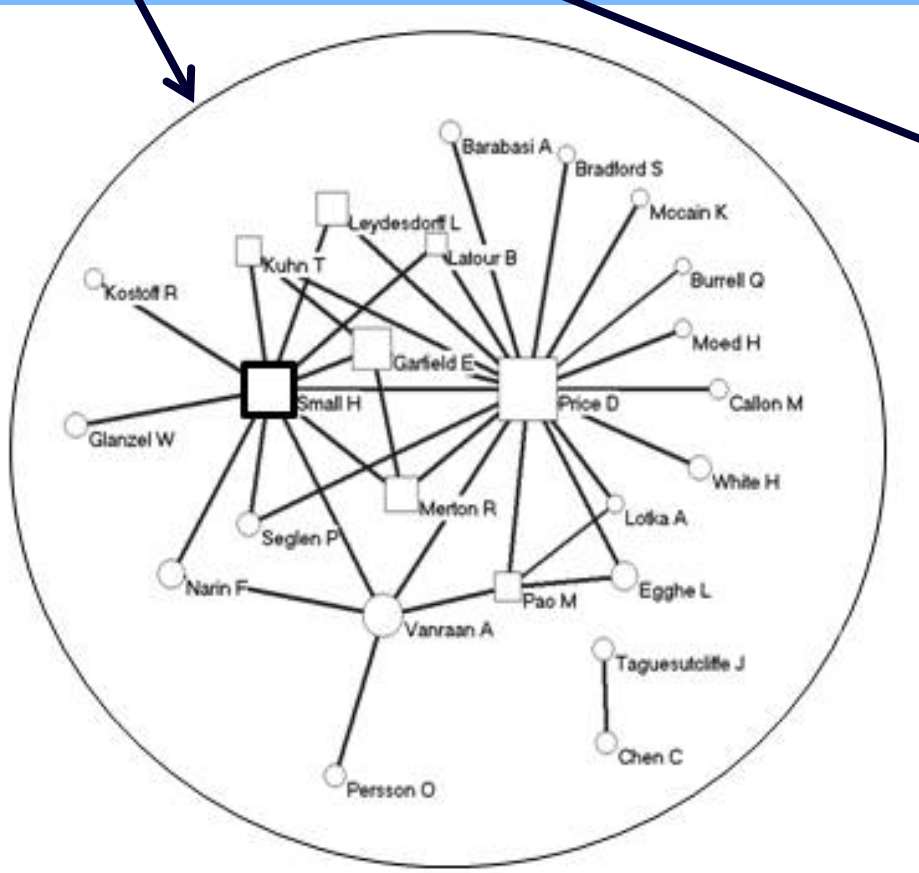


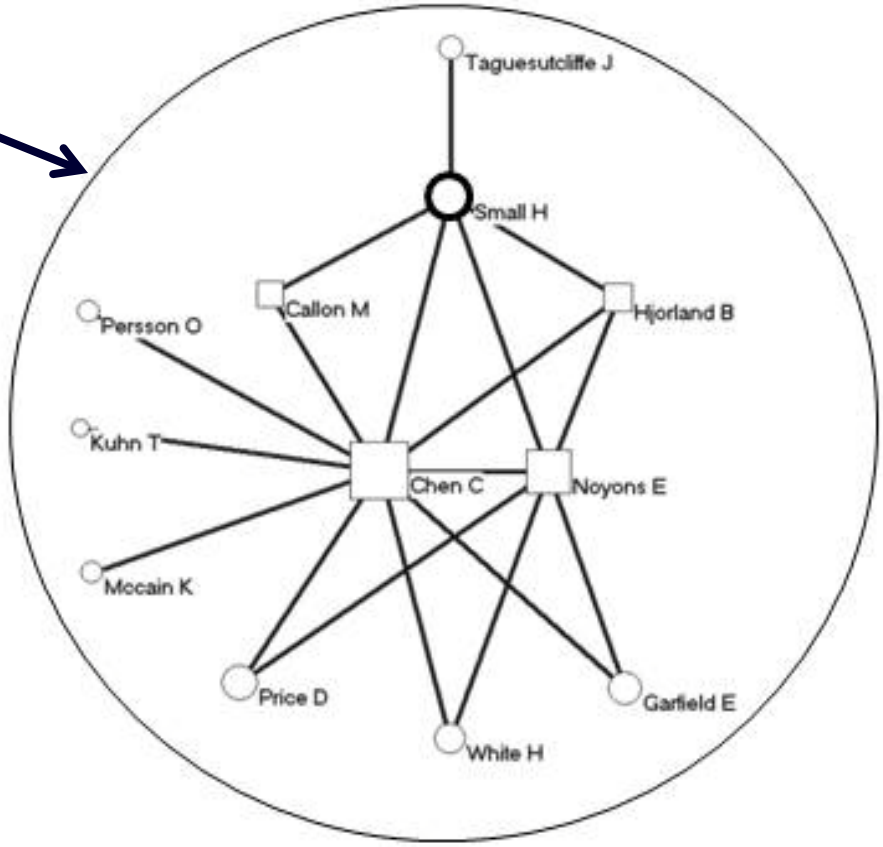


Access Innovations.

Intellectual Base of Two Research Communiities







Access Innovations.

From a society / publisher perspective visual answers

- Which topical areas form our core? Periphery?
- Where is the coverage dense? Thin?
- Which topical areas are most active? Least active?
- Which topical areas seem to be emerging? Declining?
- Which topical areas are interrelated? Isolated?
- What are the overlaps between journals / segments?
- Where are the potential expansion points?

Questions with visual answers

- From a thesaurus perspective
- What terms are too broadly defined?
- How do actual topical relationships differ from the thesaurus structure?

Access Innovations.

Preparing the data

- Index 1.2 Million Xplore records
  - Using the IEEE Thesaurus
  - Using the MeSH - Medical Subject Headings
  - Using the DTIC Thesaurus
- Normalize and enrich the XML as needed
- Create an XML / SQL Database

Access Innovations.

Mapping IEEE thesaurus space

- We are more interested in an expanded map that includes adjacencies to the IEEE data
  - Expanded term set shows adjacent white space; opportunities for expansion
  - Similar process to that for simple map except …
  - We need additional terms to add

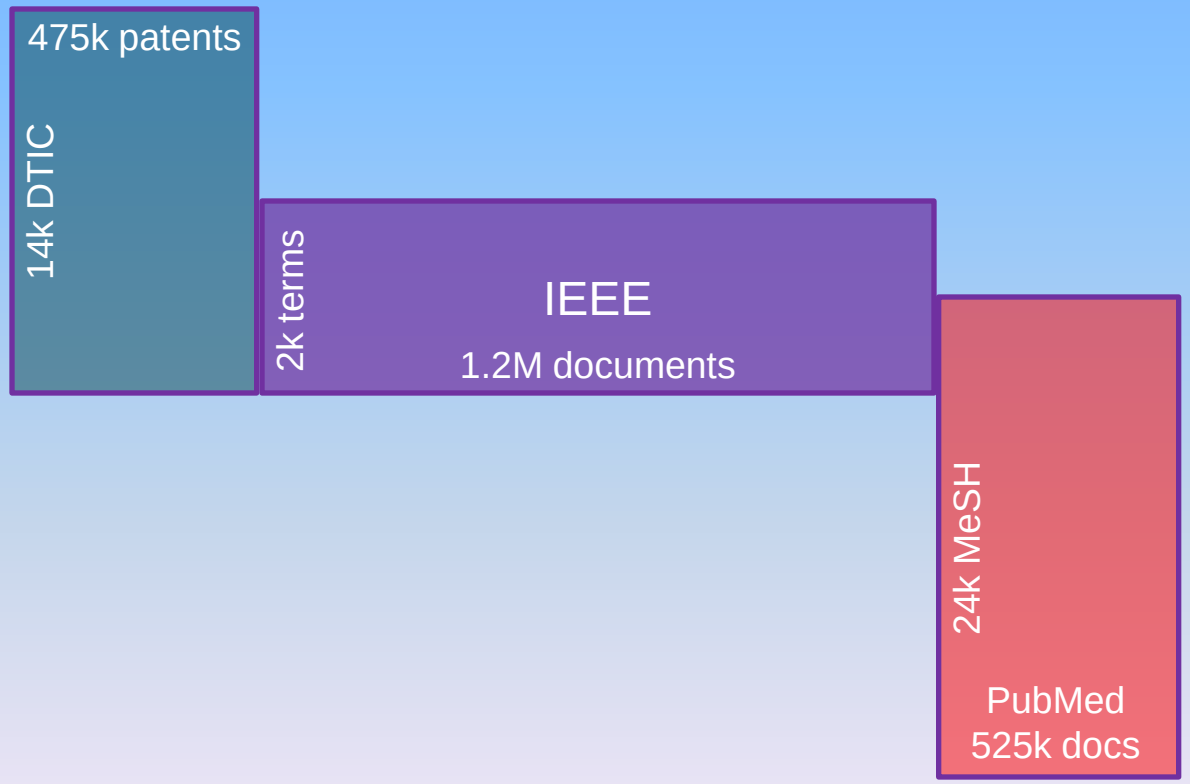
Mapping IEEE thesaurus space

- Criteria for additional terms
  - Low occurrence rate in IEEE documents
  - Linkage to terms in IEEE documents
  - Similar level of detail to current IEEE thesaurus terms
- Where do we find these terms? How can we add them?

Access Innovations.

Defining expanded term space

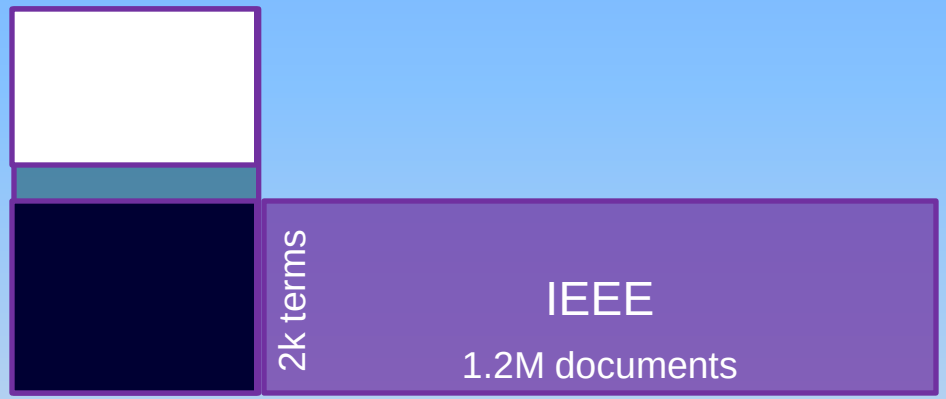
1. Select related corpus'



Access Innovations.

Defining expanded term space

2. Identify related terms



Access Innovations.

Defining expanded term space

2. Identify related terms

Access Innovations.

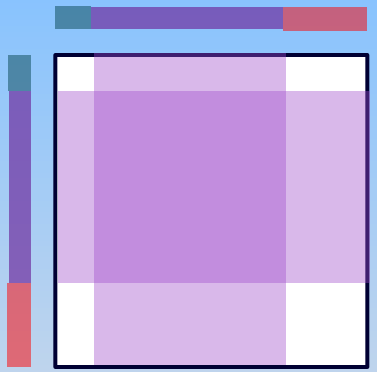
Defining expanded term space

3. Resulting term set

Access Innovations.

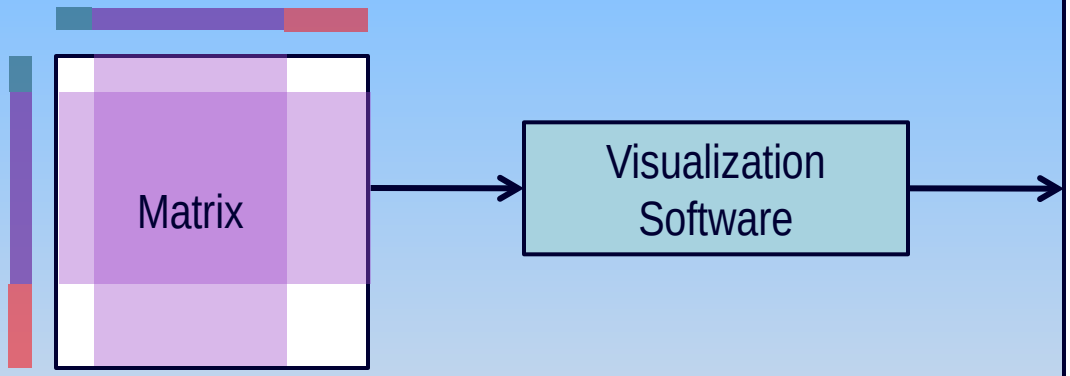
Defining expanded term space

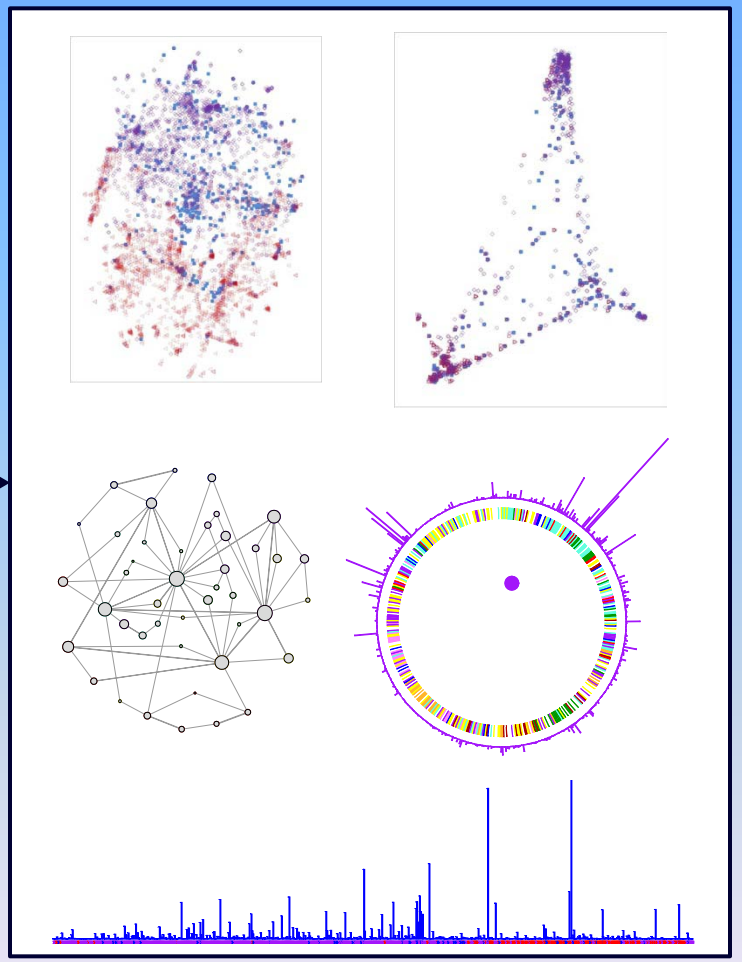
4. Term:Term Matrix



Defining expanded term space

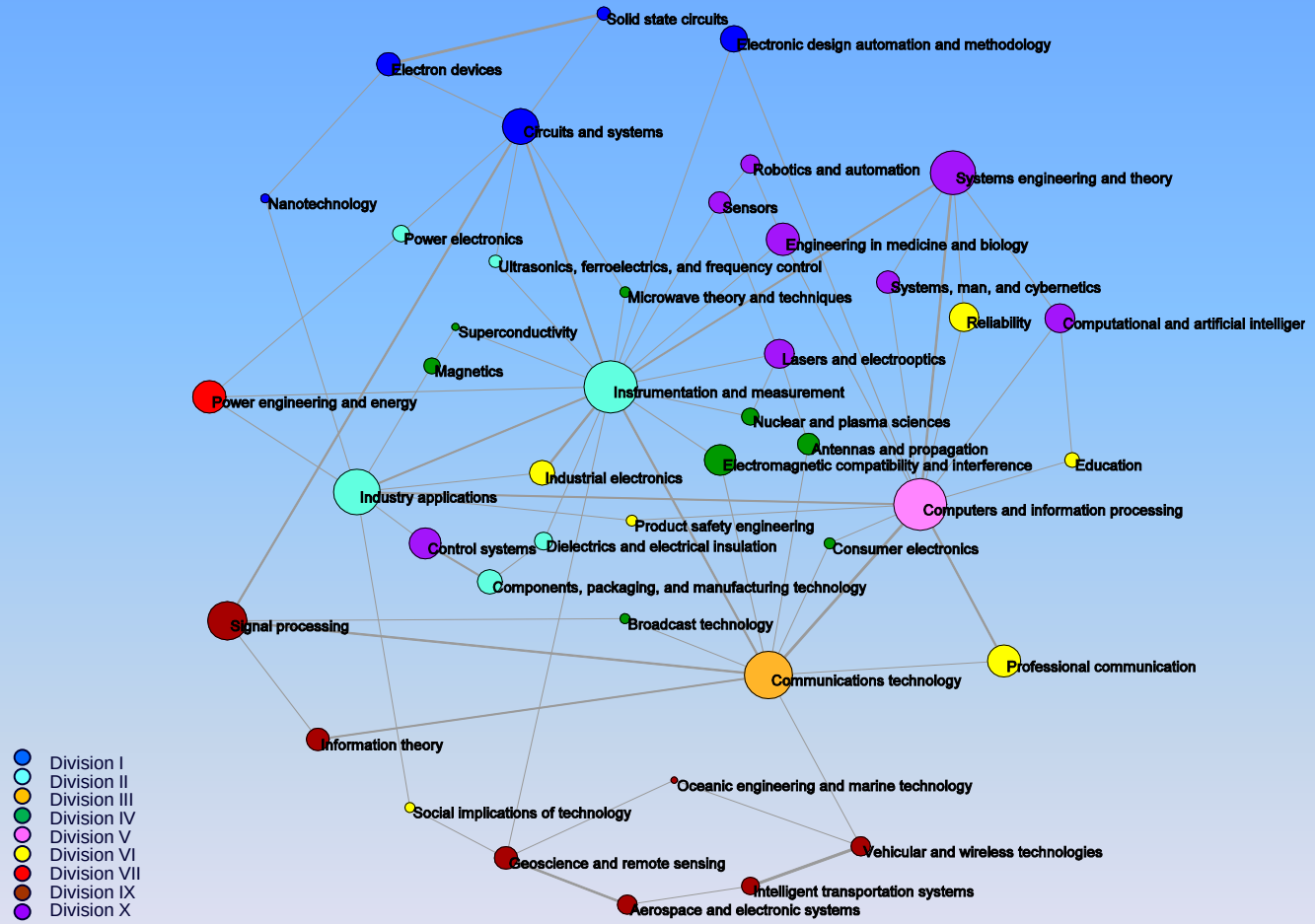
5. Visualization





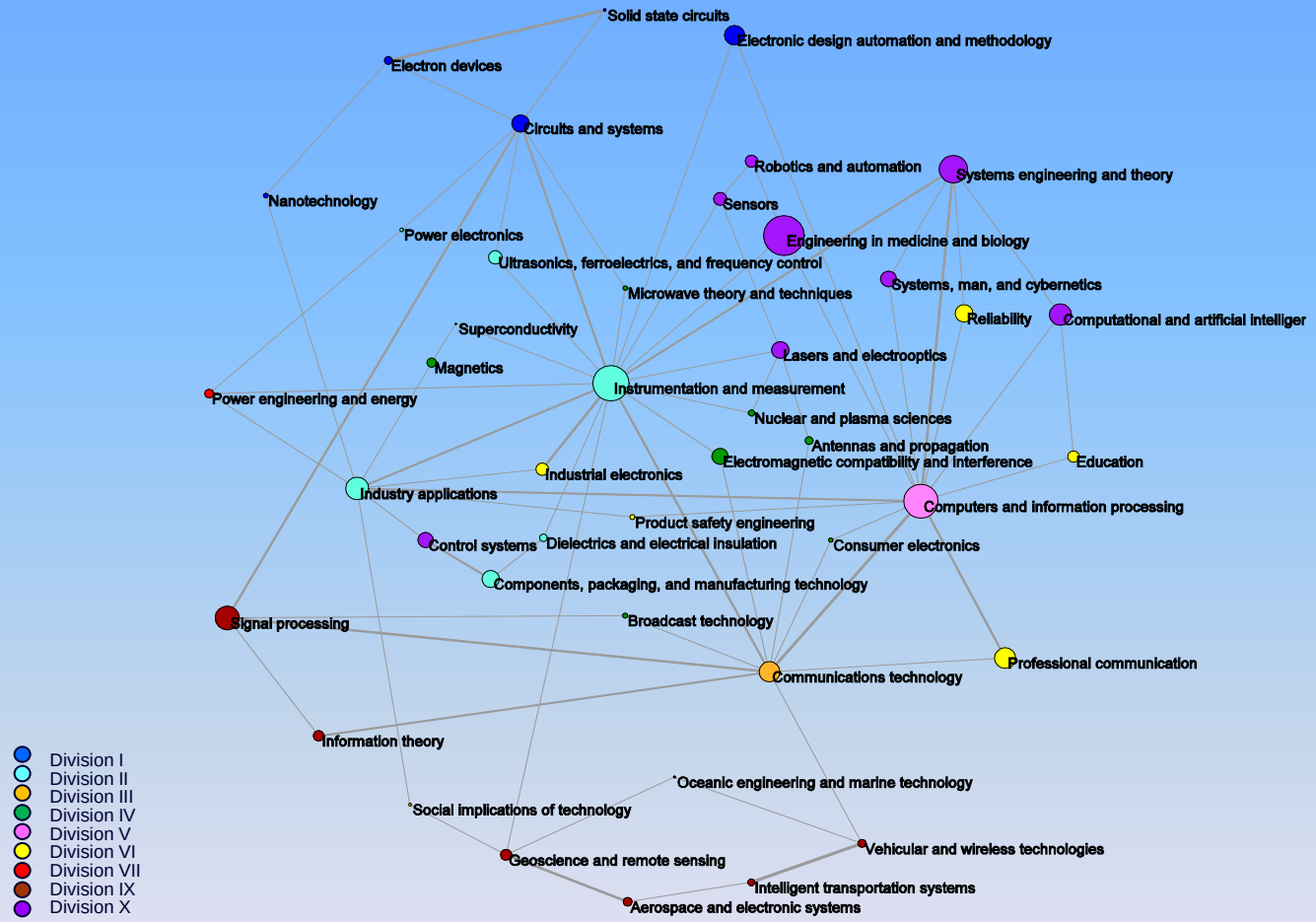
Access Innovations.

Sample term and data maps



Access Innovations.

PubMed

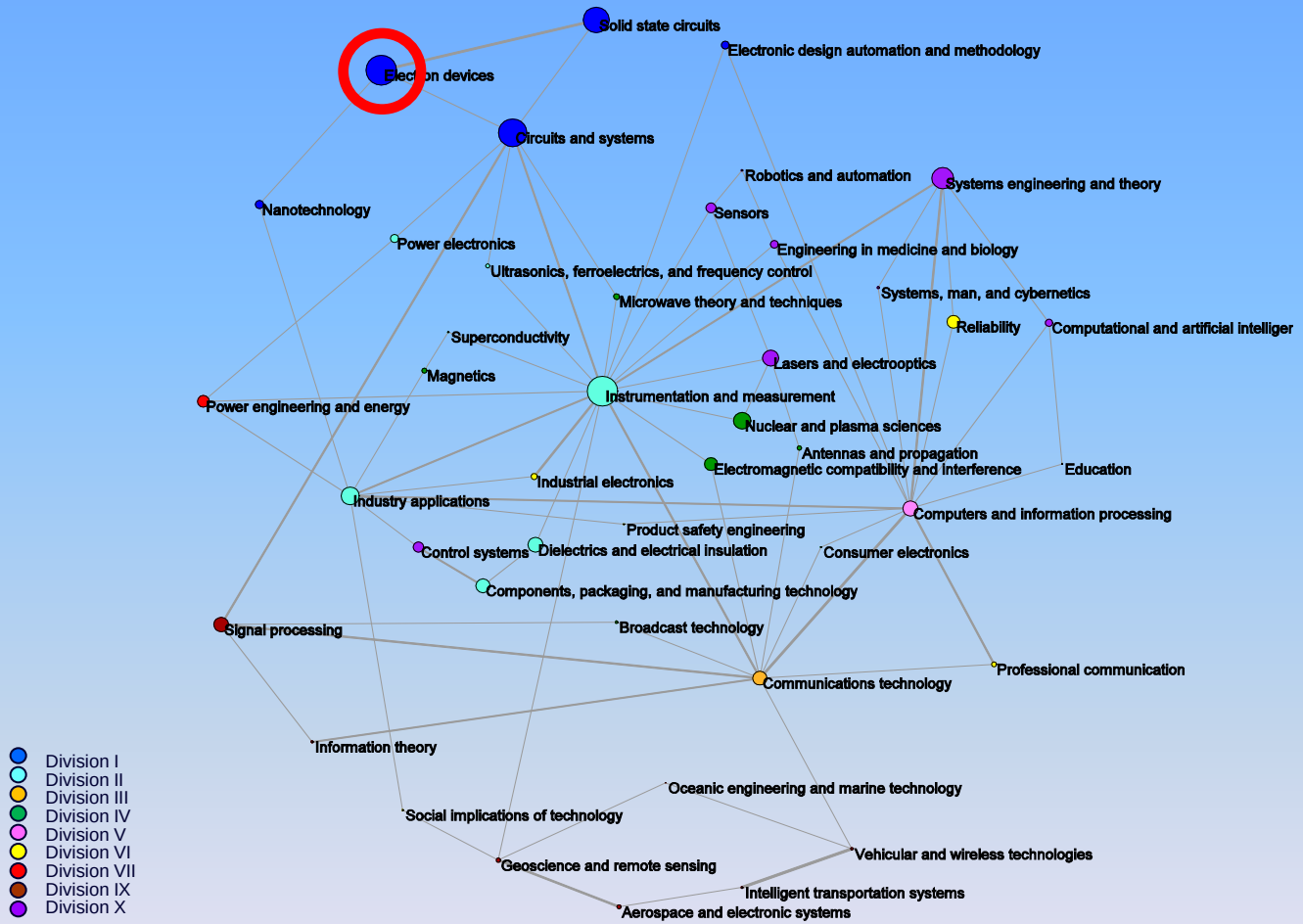


Access Innovations.

IEEE Transactions on Antennas and Propagation

Access Innovations.

IEEE Transactions on Electron Devices



Access Innovations.

Findings

- Term space can be mapped effectively
- The mapped space can be used to show distributions and trends that give answers to questions
  - Database distribution comparisons
  - Journal / segment distribution comparisons (overlaps)
  - Journal / segment trending
  - Identify groups of terms that need trimming (rule base changes)

Answering the IEEE questions

- Yes, we can use IEEE information, to forecast future directions
- Yes, look at each industry by technology sector over time to see where it is headed.
- IEEE coverage does not match the IEEE mission and vision by industry sector
- Provides new ways for IEEE to become smarter about their data and potential markets using their collection
- The societies are not all publishing and talking about what their charter indicates they cover.

Access Innovations.

Looking to the future

- We can see specific trends and which topics are emerging/cooling
- Using the IEEE data and these term analytics technology we can explore the future and the boundaries of the IEEE future

Access Innovations.

Thank You

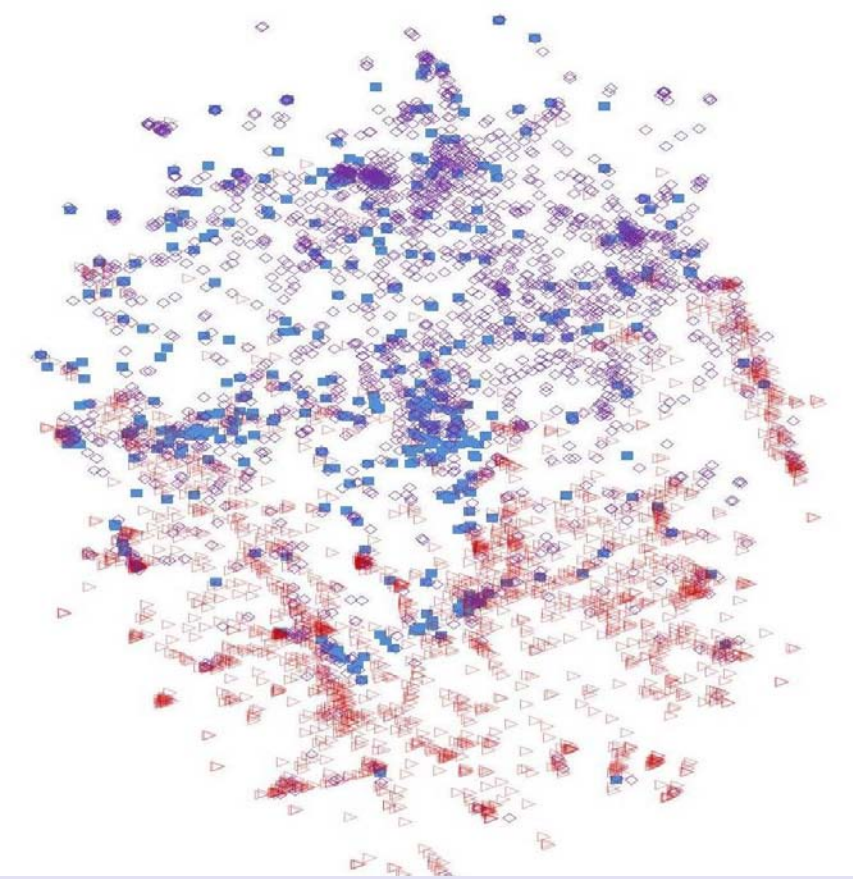
Marjorie M.K. Hlava, President,
Access Innovations / Data Harmony

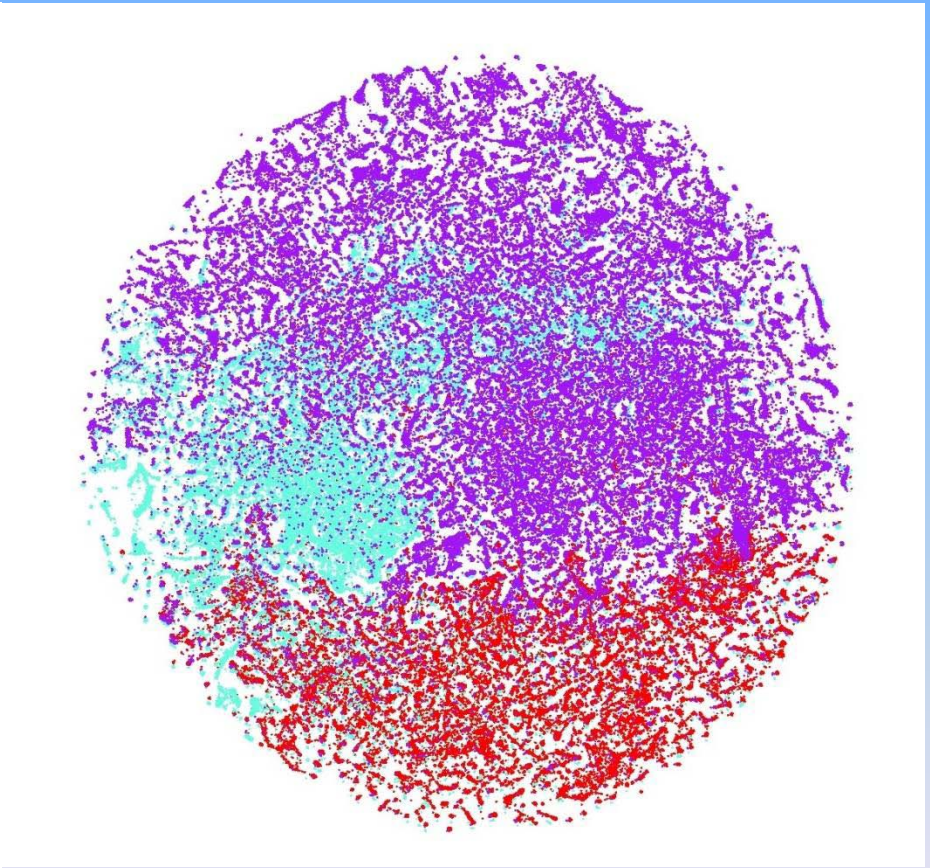
mhlava@accessinn.com

Access Innovations
4725 Indian School NE Suite 100
Albuquerque, NM 87110
www.accessinn.com
(505) 998-0800 office
(505) 256-1080 fax
www.taxodiary.com the taxonomy news blog

Access Innovations.

Consensus: Thesaurus vs. Titles

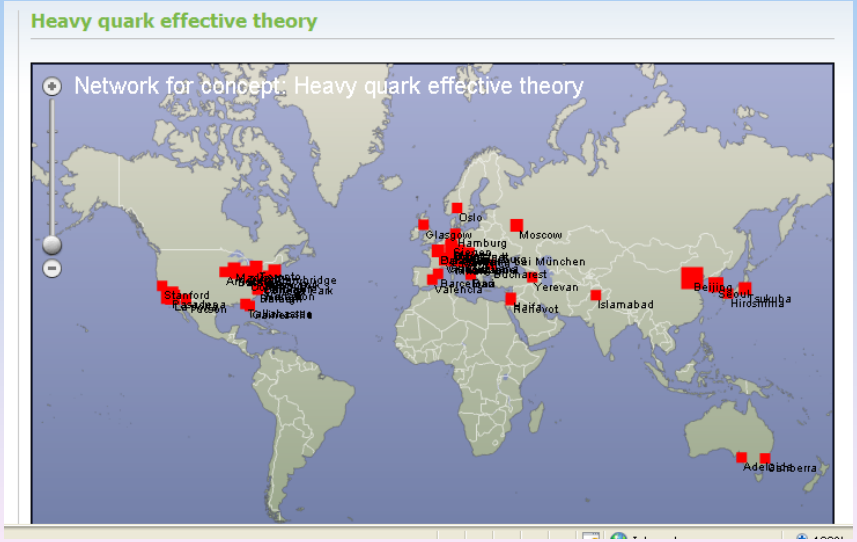




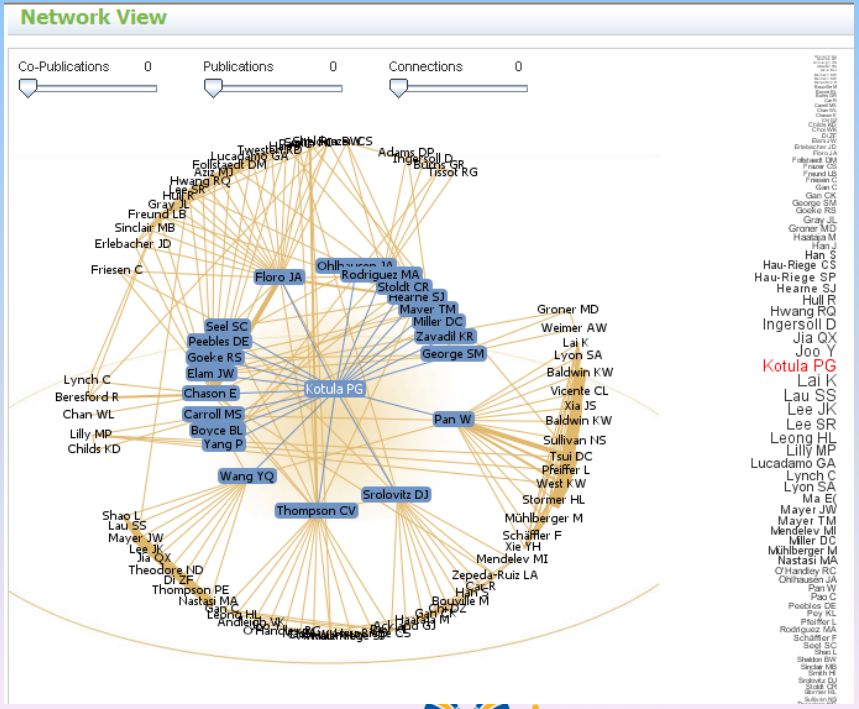
Scientific Social Networking Based on Metadata

- Idea has been here - Who is citing who like
  - ISI does it with references
  - API UniPHY does it using semantics
- Expand your options using
  - good metadata and descriptors

Map who is working in the field and where



See the authors connections



Access Innovations.