INFO116

Semantics for Enterprise Data

Enterprise Systems

- Chapter 4
- Extract, organize, and standardize (or normalize) information from many disparate and heterogeneous content sources
- For a domain of choice, identify interesting and relevant knowledge
- discover previously unknown or non-obvious relationships between documents and/or entities
- tools for fast and high-quality (contextual) querying, browsing, and analysis

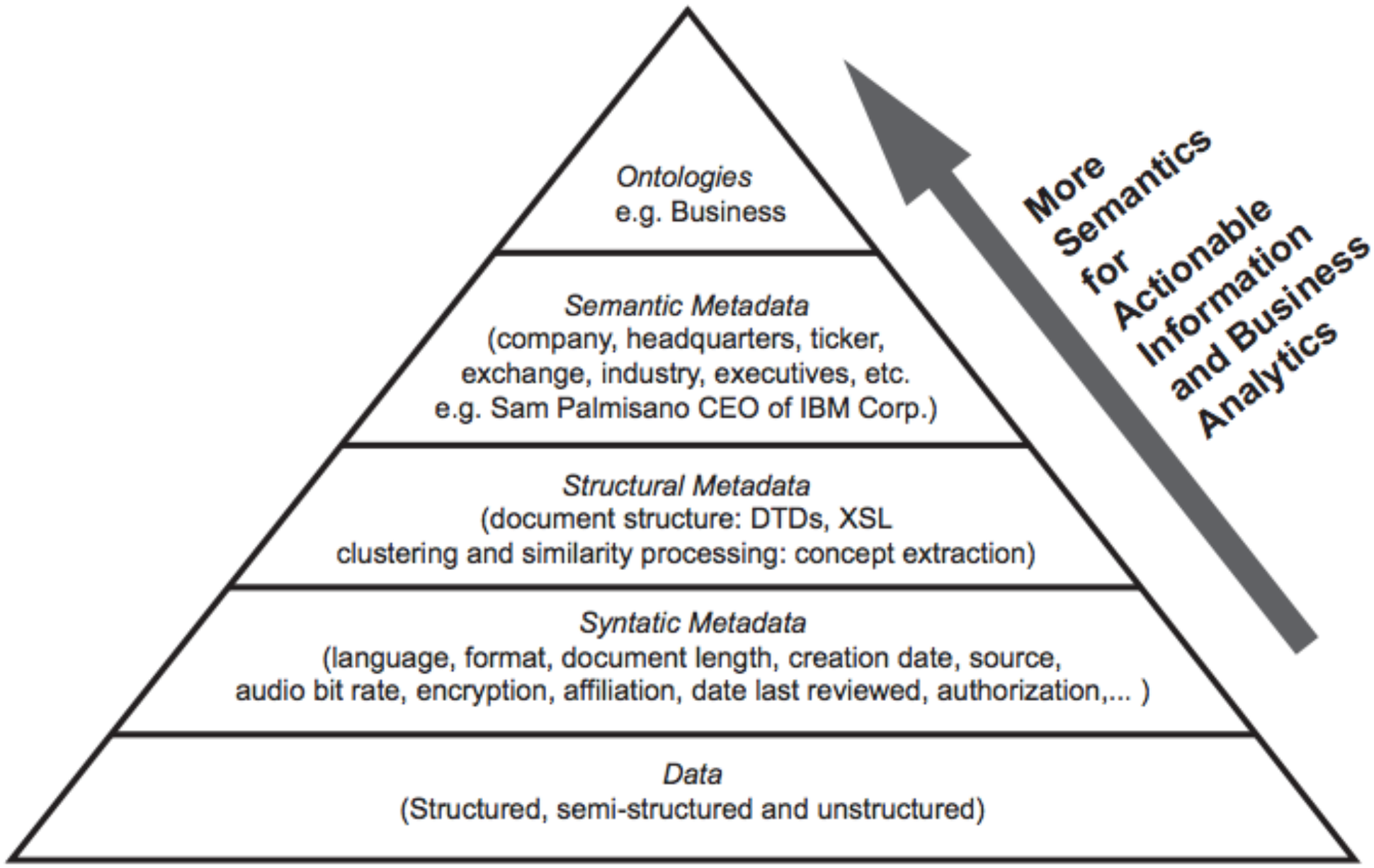
Challenges

- Excel
- Paper forms
- Departmental Information Systems
- Agility is the key to the success of modern businesses, especially in the face of globalisation
- Data is the new oil! Business analytics.

Role of semantics

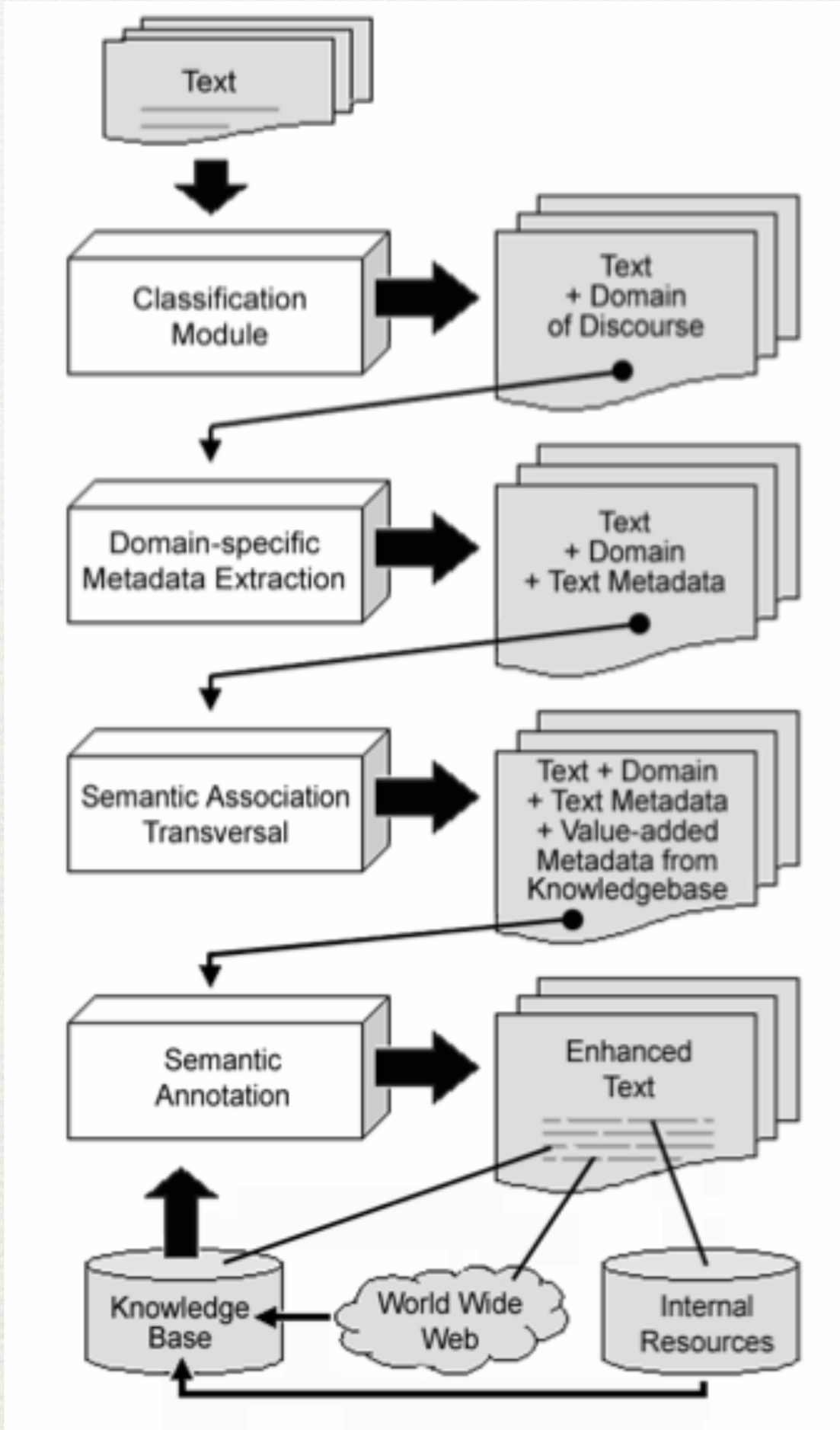
- Decision making requires comparative analysis and aggregation of content from independent sources
- Semantic organization and use of metadata (ontologies)
- Semantic normalisation (heterogeneity)
- Semantic search (synonymy, polysemy)
- Semantic association

Types of metadata

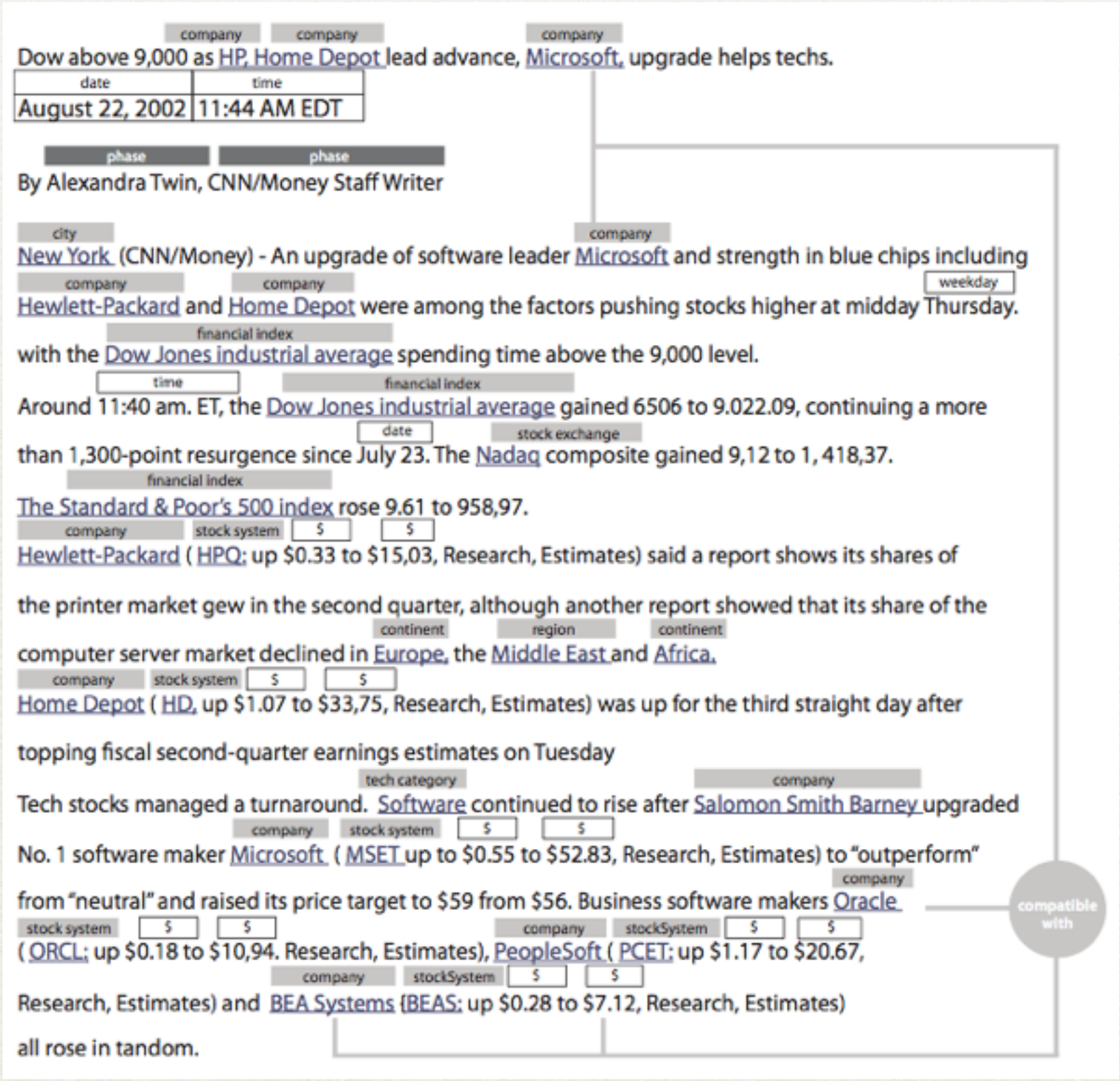


Creation of metadata

- Find, analyse, and tag relevant information
- Documents typically mention instances, not abstractions
- Dictionaries and thesauri can be used to match words and phrases to recognise and normalise *domain-specific content terms*
- Aliases and acronyms (*domain specific*)
- Documents can be analysed for various patterns and co-occurrences using extraction rules
- Ontologies, capturing *domain* (application, industry) specific knowledge (e.g. a company has a CEO, Stockholm is the capital of Sweden)


Text
Classification Module
Text + Domain of Discourse
Domain-specific Metadata Extraction
Text + Domain + Text Metadata
Semantic Association Transversal
Text + Domain + Text Metadata + Value-added Metadata from Knowledgebase
Semantic Annotation
Enhanced Text
Knowledge Base
World Wide Web
Internal Resources

Semantic Enhancement



General approach

- Creation of a schema that serves as the definitional component of the ontology
- Population of the ontology at the instance level
- Metadata extraction or semantic annotation of heterogeneous content from a variety of sources
- Blended Semantic Browsing and Querying (BSBQ) of content to let user seamlessly cross-navigate between related knowledge and content

SCORE

- Semantic Content Organization and Retrieval Engine
- SCORE Enhancement Engine
  - different modules for heterogeneous document types
  - reliable automatic classification of documents
  - accurate extraction of semantic, domain-specific metadata
  - extensive management of the enhancement processes including reporting and semantic annotation mechanisms

SCORE

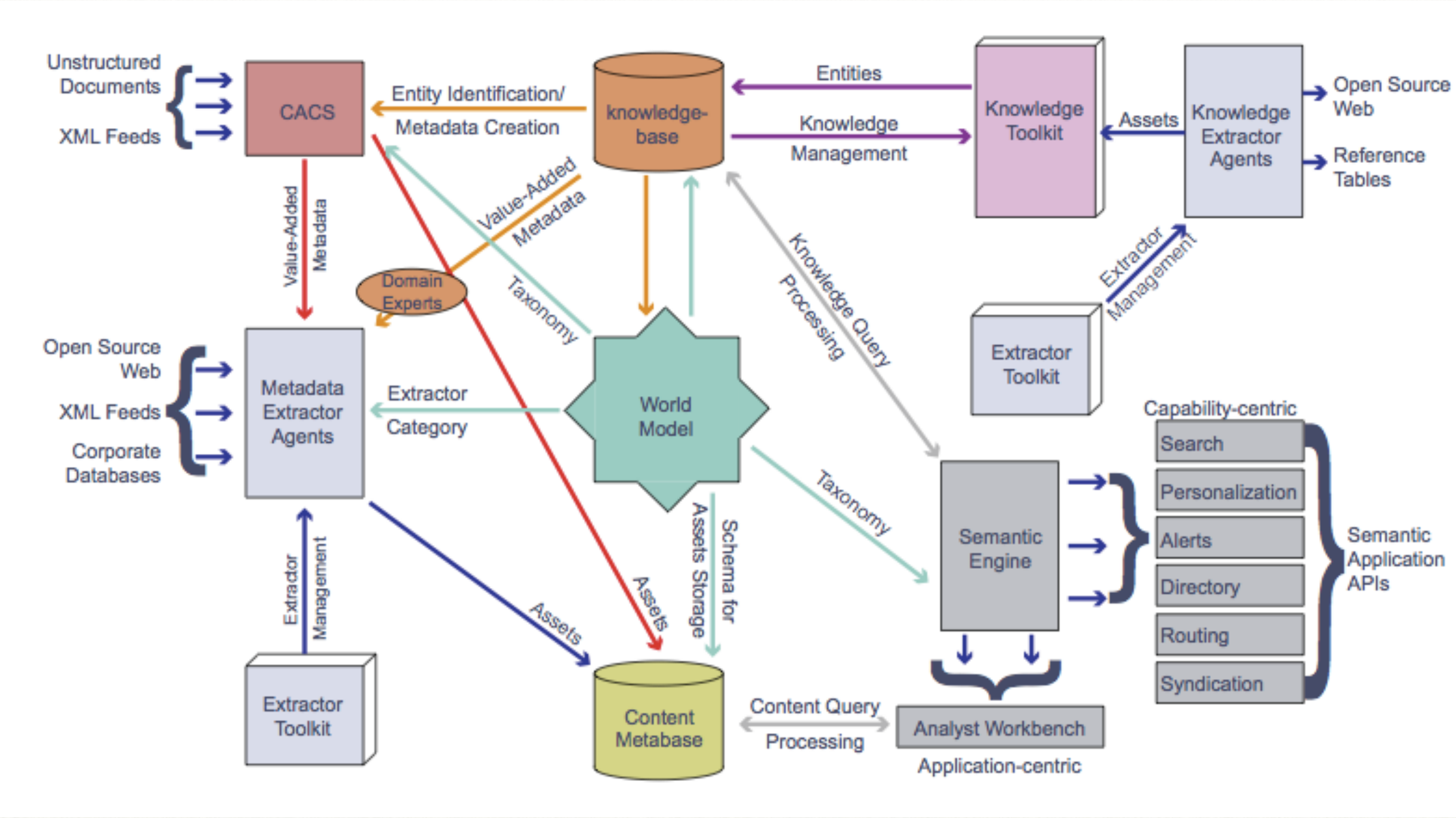
- World Model contains definitions
- Knowledgebase contains assertions
  - named entities (people, places)

Knowledgebase

Oracle Corp.

Sector:	Computer Software and Services
Industry:	Database and File Management Software
Symbol:	ORCL
CEO:	Ellison, Lawrence J.
CFO:	Henley, Jeffrey O.
Headquartered in:	RedWood City, California, USA
Manufactured by:	8i Standard Edition, Application Server, etc.
Subsidiary of:	Liberate Technologies and OracleMobile
Competes with:	Agile, Ariba, BEA Systems, Informix, IBM, Microsoft, PeopleSoft and Sybase

Architecture



Enabling Semantic Tasks

- text classification
- metadata extraction
- identification of semantic associations
- enrichment of a document by annotating it with pertinent semantic information

“Knowledge Graph”

- When Entities are discovered in a document, other Entities can be attached to the document through the traversal of semantic associations. These traversals can be single-step (for example “Linux” has “created by” relationship with “Linus Torvalds”) or may follow user-defined paths (such as “Microsoft” has a “competes with” relationship with “Sun Microsystems” which has a “created” relationship with “Java.”) This allows for connections to be created between entries in the Metabase by exploiting domain-specific knowledge. This represents a unique functionality of SCORE.

Linking Enterprise Data

- Escape the data silo!
- LED (Linked Enterprise Data)
- information creation is intimately coupled with the act of information sharing

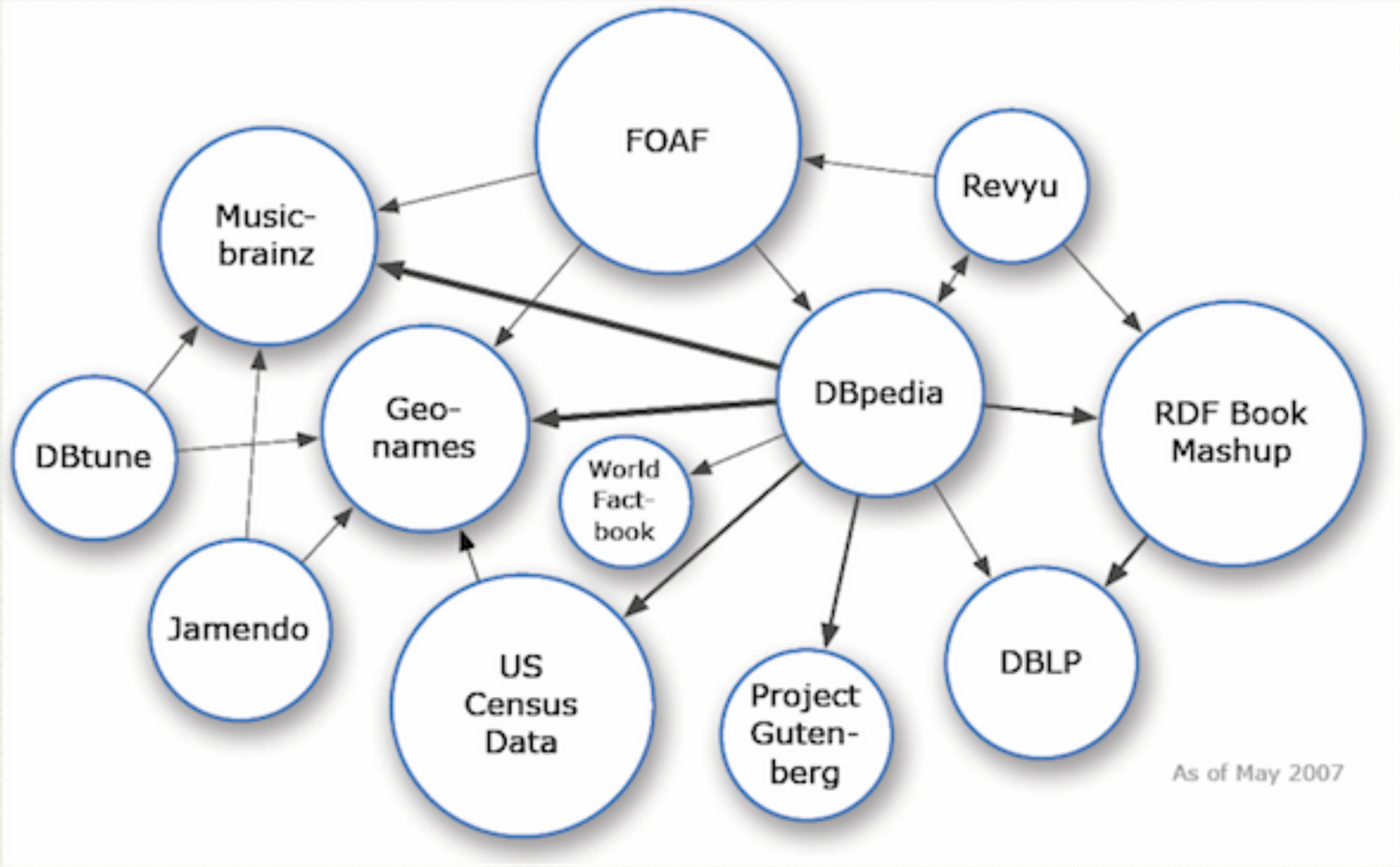
Linked Data

- Linked Data is about using the Web to connect related data that wasn't previously linked, or using the Web to lower the barriers to linking data currently linked using other methods.

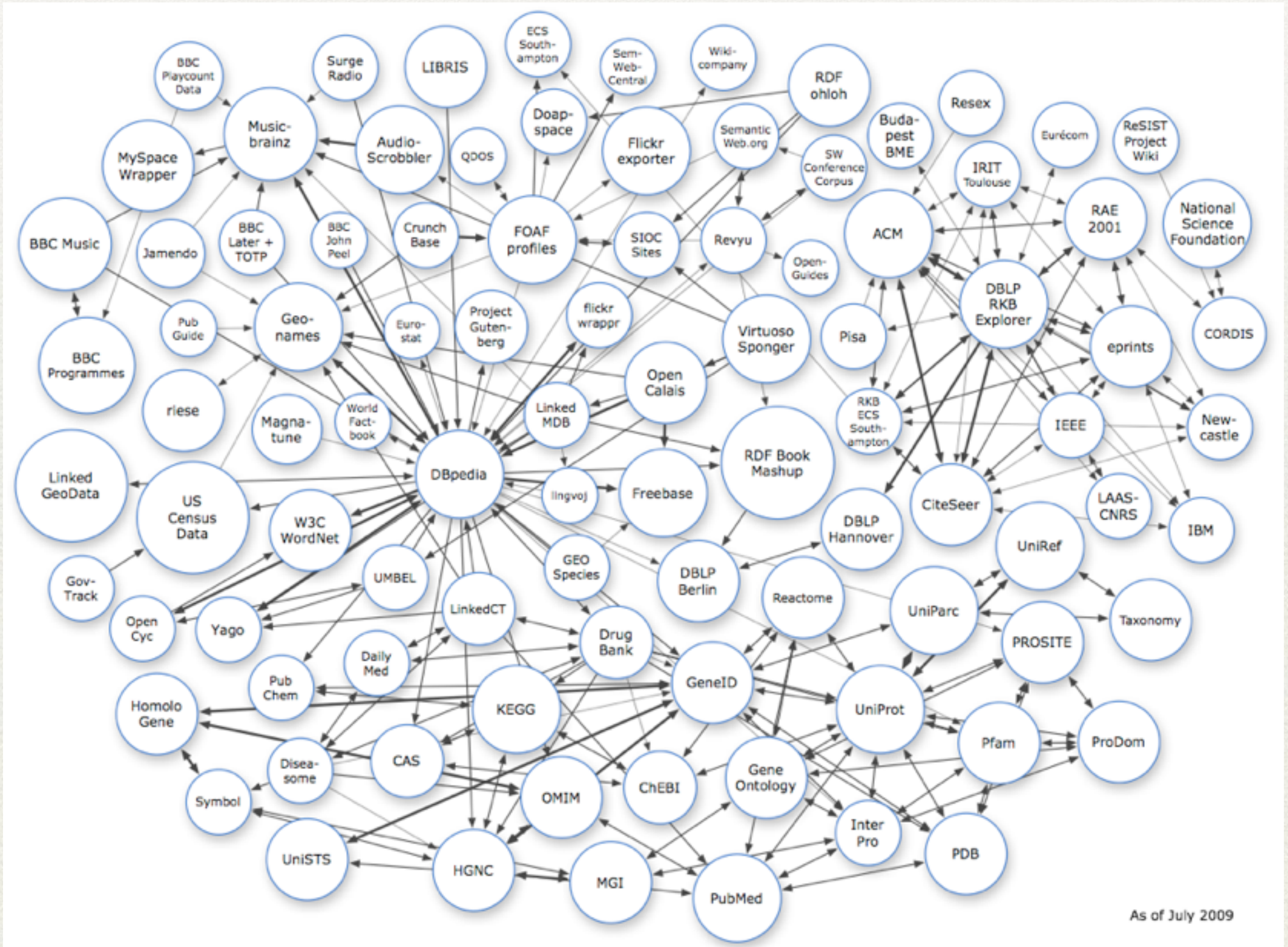
Linked Data Design

- TBL http://www.w3.org/DesignIssues/LinkedData.html
  - Use URIs as names for things
  - Use HTTP URIs so that people can look up those names.
  - When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
  - Include links to other URIs. so that they can discover more things.

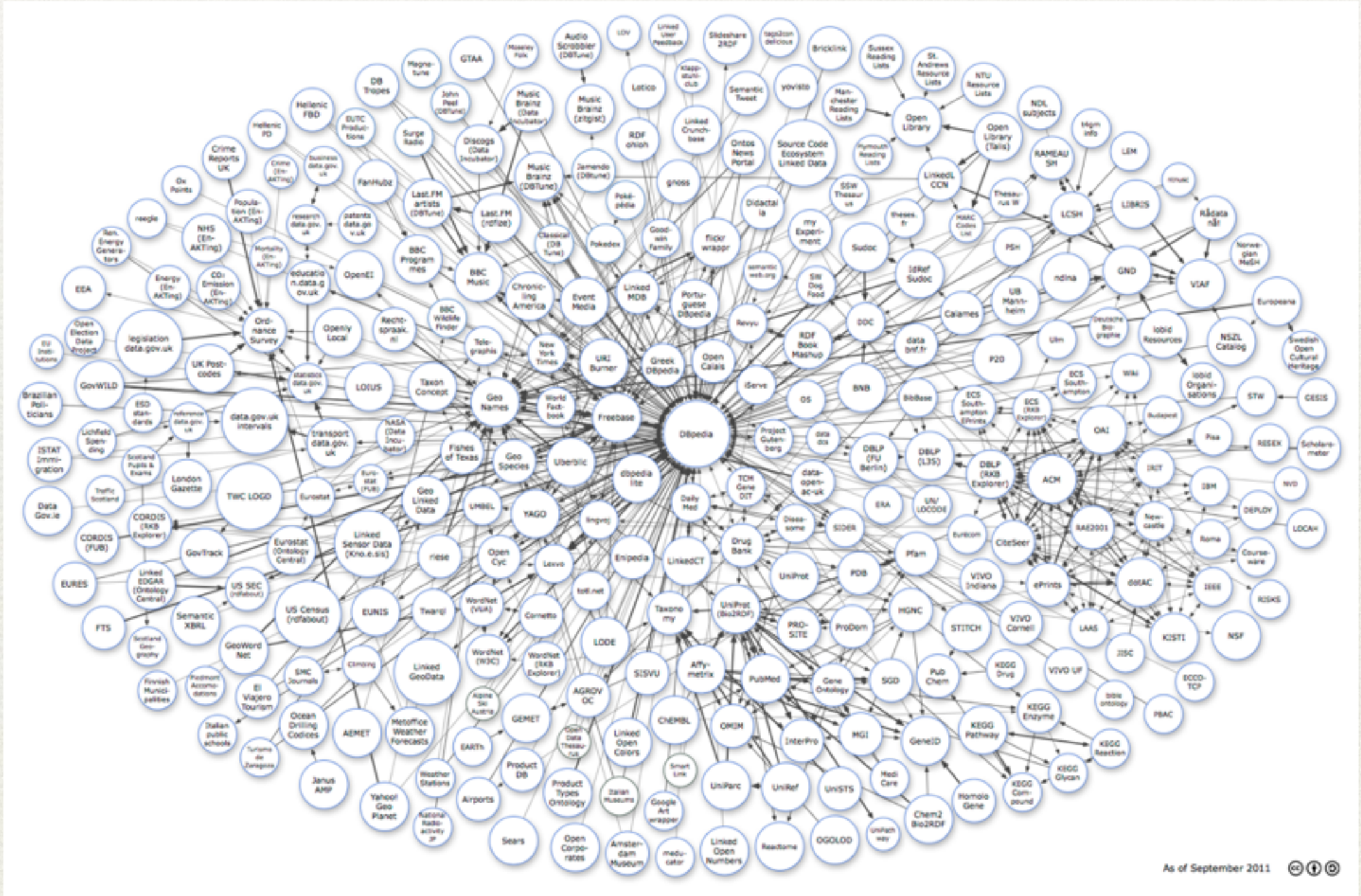
LOD Cloud 2007-05-01



LOD Cloud 2009-03-05



LOD Cloud 2011-09-19



2014

- http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/

Datasets by topical domain.

Topic	Datasets	%
Government	183	18.05%
Publications	96	9.47%
Life sciences	83	8.19%
User-generated content	48	4.73%
Cross-domain	41	4.04%
Media	22	2.17%
Geographic	21	2.07%
Social web	520	51.28%
Total	1014	

5 Star data

★ Available on the web (whatever format) *but with an open licence, to be Open Data*

★★ Available as machine-readable structured data (e.g. excel instead of image scan of a table)

★★★ as (2) plus non-proprietary format (e.g. CSV instead of excel)

★★★★ All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff

★★★★★ All the above, plus: Link your data to other people's data to provide context

Provenance

- Data: where did it come from? How did it get there?
  - Can it be trusted?
  - Machine interpretable.

Enterprise Provenance

- Companies are subject to a wide range of rules and regulations that apply directly to their internal operations and the products and services they provide. Compliance to these regulations are essential for a company to operate transparently and ethically in their particular markets. They are required to prove compliance to the imposed regulations through internal and external auditing processes .

Provenance dimensions

Category	Dimension	Description
Content	Object	The artifact that a provenance statement is about.
	Attribution	The sources or entities that contributed to create the artifact in question.
	Process	The activities (or steps) that were carried out to generate or access the artifact at hand.
	Versioning	Records of changes to an artifact over time and what entities and processes were associated with those changes.
	Justification	Documentation recording why and how a particular decision is made.
	Entailment	Explanations showing how facts were derived from other facts.
Management	Publication	Making provenance available on the Web.
	Access	The ability to find the provenance for a particular artifact.
	Dissemination	Defining how provenance should be distributed and its access be controlled.
	Scale	Dealing with large amounts of provenance.
Use	Understanding	How to enable the end user consumption of provenance.
	Interoperability	Combining provenance produced by multiple different systems.
	Comparison	Comparing artifacts through their provenance.
	Accountability	Using provenance to assign credit or blame.
	Trust	Using provenance to make trust judgments.
	Imperfections	Dealing with imperfections in provenance records.
	Debugging	Using provenance to detect bugs or failures of processes.

PROV-O: The PROV Ontology

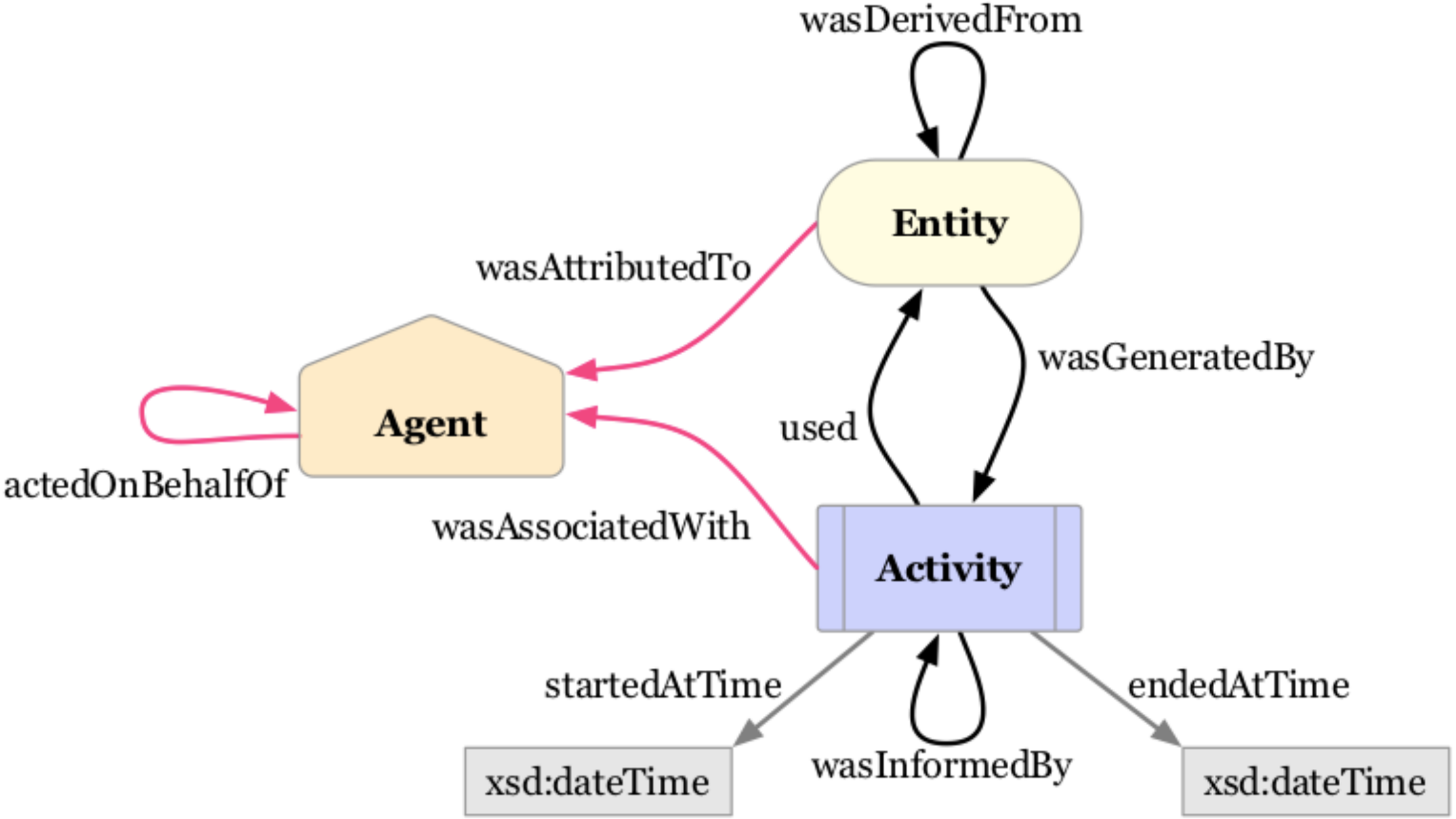


Figure 1. The three Starting Point classes and the properties that relate them.
The diagrams in this document depict Entities as yellow ovals,
Activities as blue rectangles, and Agents as orange pentagons.
The responsibility properties are shown in pink.

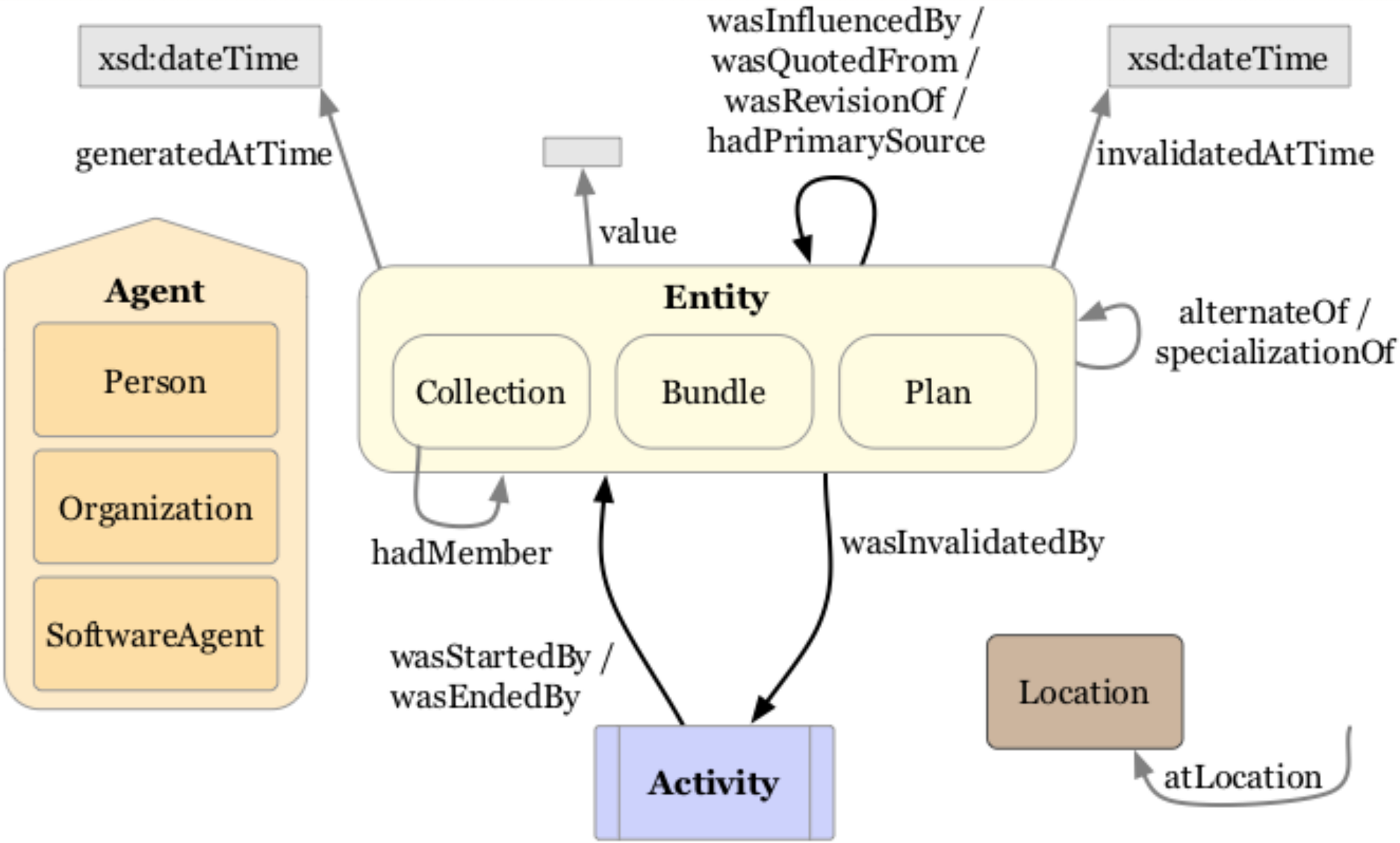


Figure 3. The expanded terms build upon those in the Starting Points section.
The diagrams in this document depict Entities as yellow ovals, Activities as blue rectangles, and Agents as orange pentagons.
The domain of prov:atLocation (`prov:Activity` or `prov:Entity` or `prov:Agent` or `prov:InstantaneousEvent`) is not illustrated.

Oracle Press Release

Oracle Implements W3C's Standard for Data Provenance

Standardization Enables Organizations to Quickly and Easily Integrate Advanced Control Solutions Across Heterogeneous Systems

Redwood Shores, Calif. – October 4, 2013

News Summary

The growth of data is well documented, but it is easy to forget that there are a range of people, entities and activities involved in producing each electronic record. For organizations this additional level of complexity can present serious operational risk and process effectiveness challenges. Without a standard way of assessing the provenance of electronic records, it is extremely difficult for organizations to determine the quality of information, accurately attribute it to the correct author(s) or prove that actions have been performed according to specification. It is also extremely challenging to put the proper controls in place without provenance of information in heterogeneous environments. To address these challenges and provide customers with complete and virtually seamless controls coverage, Oracle has implemented the World Wide Web (W3C) Provenance (PROV) standard in the latest release of Oracle Fusion Advanced Controls.

Example

```
@prefix xsd:  <http://www.w3.org/2001/XMLSchema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix :     <http://example.org#> .

:bar_chart
   a prov:Entity;
   prov:wasGeneratedBy  :illustrationActivity;
   prov:wasDerivedFrom  :aggregatedByRegions;
   prov:wasAttributedTo :derek;
.

:derek
   a foaf:Person, prov:Agent;
   foaf:givenName       "Derek";
   foaf:mbox            <mailto:derek@example.org>;
   prov:actedOnBehalfOf :natonal_newspaper_inc;
.

:national_newspaper_inc
   a foaf:Organization, prov:Agent;
   foaf:name "National Newspaper, Inc.";
.

:illustrationActivity
   a prov:Activity;
   prov:used              :aggregatedByRegions;
   prov:wasAssociatedWith :derek;
   prov:wasInformedBy     :aggregationActivity;
.

:aggregatedByRegions
   a prov:Entity;
   prov:wasGeneratedBy  :aggregationActivity;
   prov:wasAttributedTo :derek;
.

:aggregationActivity
   a prov:Activity;
   prov:startedAtTime     "2011-07-14T01:01:01Z"^^xsd:dateTime;
   prov:wasAssociatedWith :derek;
   prov:used               :crimeData;
   prov:used               :nationalRegionsList;
   prov:endedAtTime       "2011-07-14T02:02:02Z"^^xsd:dateTime;
.

:crimeData
   a prov:Entity;
   prov:wasAttributedTo :government;
.
:government a foaf:Organization, prov:Agent .

:nationalRegionsList
   a prov:Entity;
   prov:wasAttributedTo :civil_action_group;
.
:civil_action_group a foaf:Organization, prov:Agent .
```