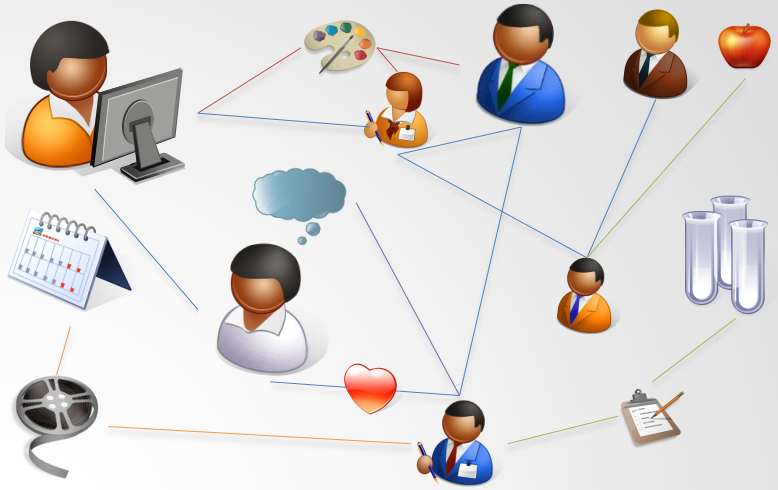


ZENTITY

A repository and a platform for building repositories

Investigations into “Semantic Computing”

Savas Parastatidis
Software Architect/Philosopher
Bing (previously with MSR)

blog: http://savas.me
twitter: http://twitter.com/savasp
facebook: http://facebook.com/savas
email: savas@parastatidis.name

AGENDA

Introductions

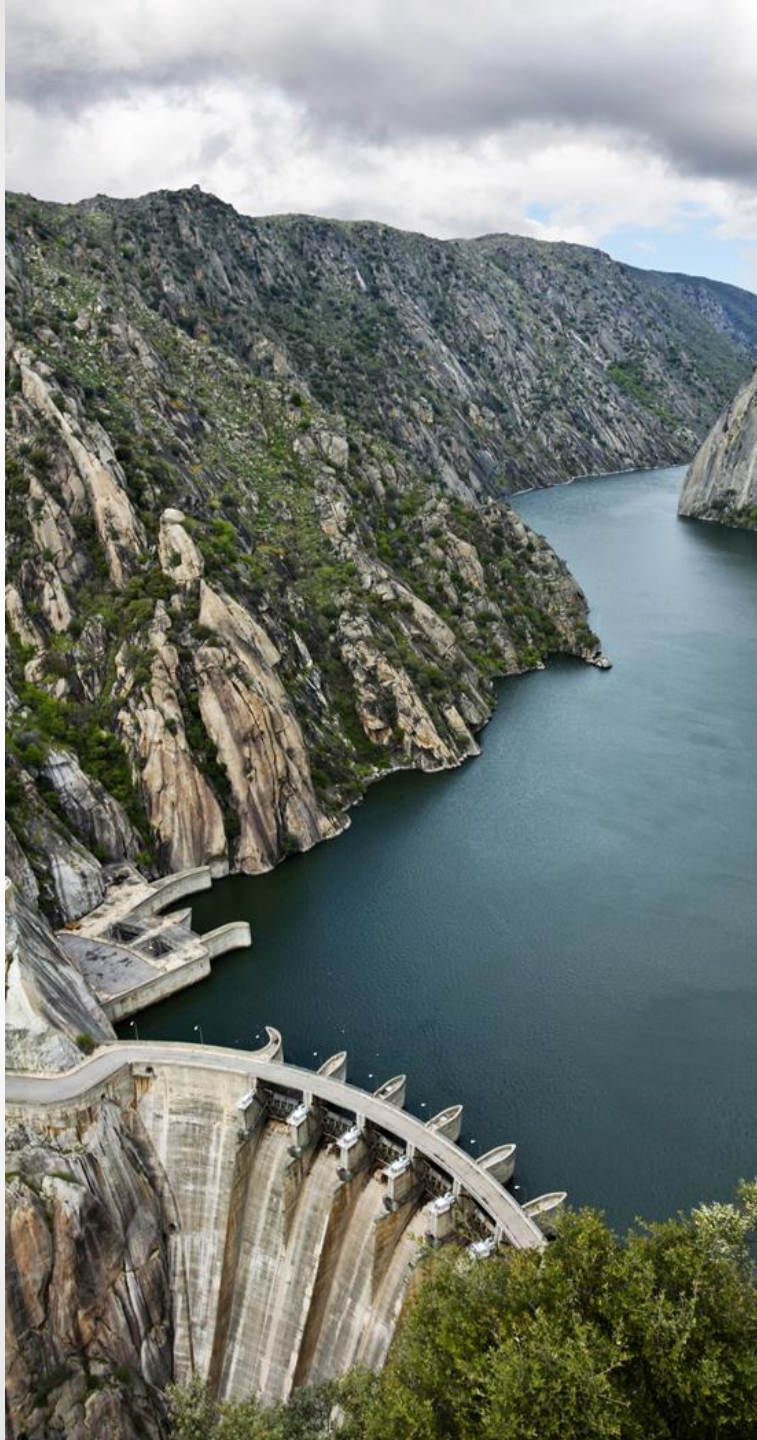
Context of Research Repositories

Subjective view of Semantic Computing concepts

Zentity Architecture

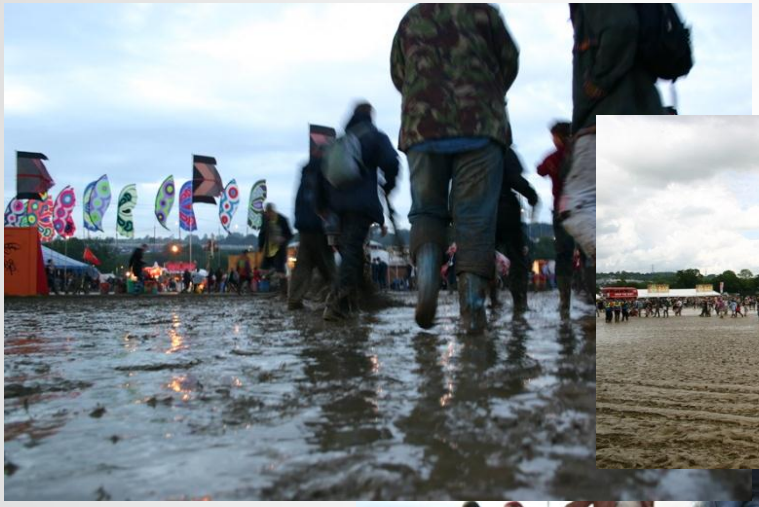
Samples/Demos

But I love to improvise so please ask questions along the way!



DO ASK QUESTIONS

Unfortunately, I have to leave tonight to go to this…



WHO ARE WE?

Microsoft Research

External Research
Tony Hey (Corporate Vice President)
Daron Green (Senior Director)

Education and Scholarly Communications
Lee Dirks (Director)
Alex Wade (Director)
Pablo Fernicola (Group Program Manager)

MISSION

Optimize and extend Microsoft software to meet the specific needs of the academic community

Our approach:

Conduct applied projects to enhance academic productivity by evolving Microsoft's scholarly communication offerings

Microsoft External Research is uniquely positioned to drive this initiative across Microsoft

Community and Geographic Outreach

Core Computer Science

Earth, Energy and Environment

Education & Scholarly Communications

Health & Wellbeing

Advanced Research Tools and Services

WHO WE WORK WITH

Academics / Scholars

Researchers / Scientists

Libraries / Archives

Scholarly Publishers & Societies

Governments / Policy Organizations

OUR BUSINESS MODEL?

Dual benefit

WHO AM I?

WHO AM I?

Background

Parallel/distributed computing

eScience

Web Services

Web

Grid

Cloud

Semantic Computing

WHO AM I?

Now with Bing (doesn't it look good? :-) to work on discovery of information using semantics

Previously an Architect in External Research

Responsible for projects like:
Ontology plugin for Word
Chem4Word
Creative Commons plugin

and of course

Zentity

RESEARCH OUTPUT REPOSITORIES

Journal articles

☑ peer-review, indexed, archived

☒ timeliness, cost, access, format limits

Digital Repositories

Subject Repositories

arXiv.org (Physics, Math, CS)

PubMed Central (Biomedical)

Institutional Repositories

Data repositories

Data sets, presentations, workflows, etc.

Research & Analysis

- Research Information Center
- Researcher Desktop

Authoring

- Article Authoring Add-in for Word
- Creative Commons Office Add-in
- Semantic Annotations in Word
- Reproducible Research in Word

Publication & Dissemination

- eJournal Service
- Conference Management Tool

Storage & Archiving

- Research Output Repository

SEMANTIC COMPUTING

DOES NOT IMPLY

RDF, OWL, ETC.

SEMANTICS

Term used to refer to the concept of “meaning”

The linguistics, AI, Natural Language Processing, etc. communities have been working on “meaning” and “knowledge” related technologies for decades

Emergence of a new breed of technologies to capture meaning (RDF, OWL, etc.)

WHAT IS SEMANTIC COMPUTING?

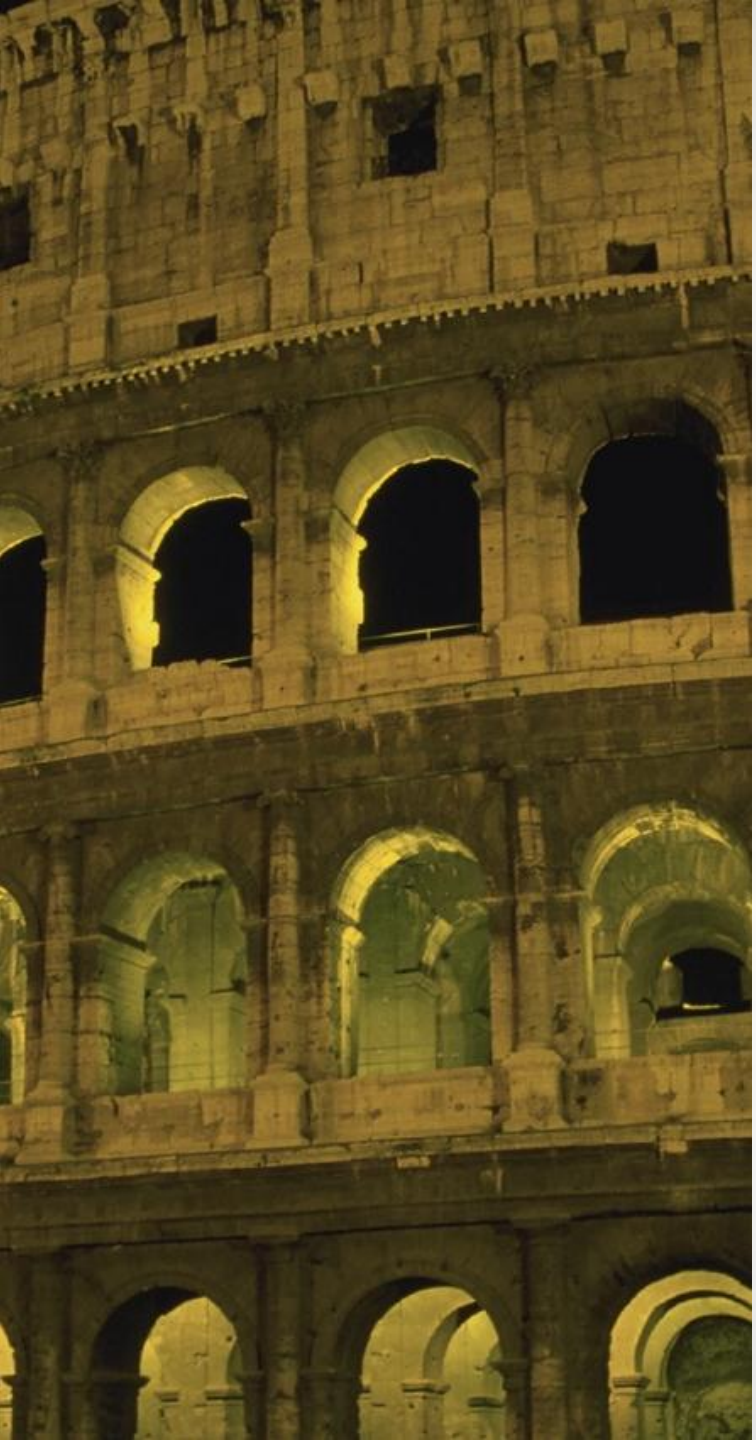
Set of concepts/technologies/approaches

Data modeling, structured data, relationships, ontologies

Machine learning (entity extraction)

Inference, reasoning

…



SET OF TECHNOLOGIES TO...

Model data and their connections
e.g. RDF, Topic Maps, Unified Content Descriptors

Capture concepts and their relationships
e.g. OWL

Query data and produce information
e.g. SPARQL

Reason about data, concepts, information
e.g. Pellet

Extract structured information (machine learning)
e.g. Live Labs entity extraction

TODAY…

Computers are great **tools** in

huge amounts of **data**

For example, Google and Microsoft both have copies of the Web for indexing purposes

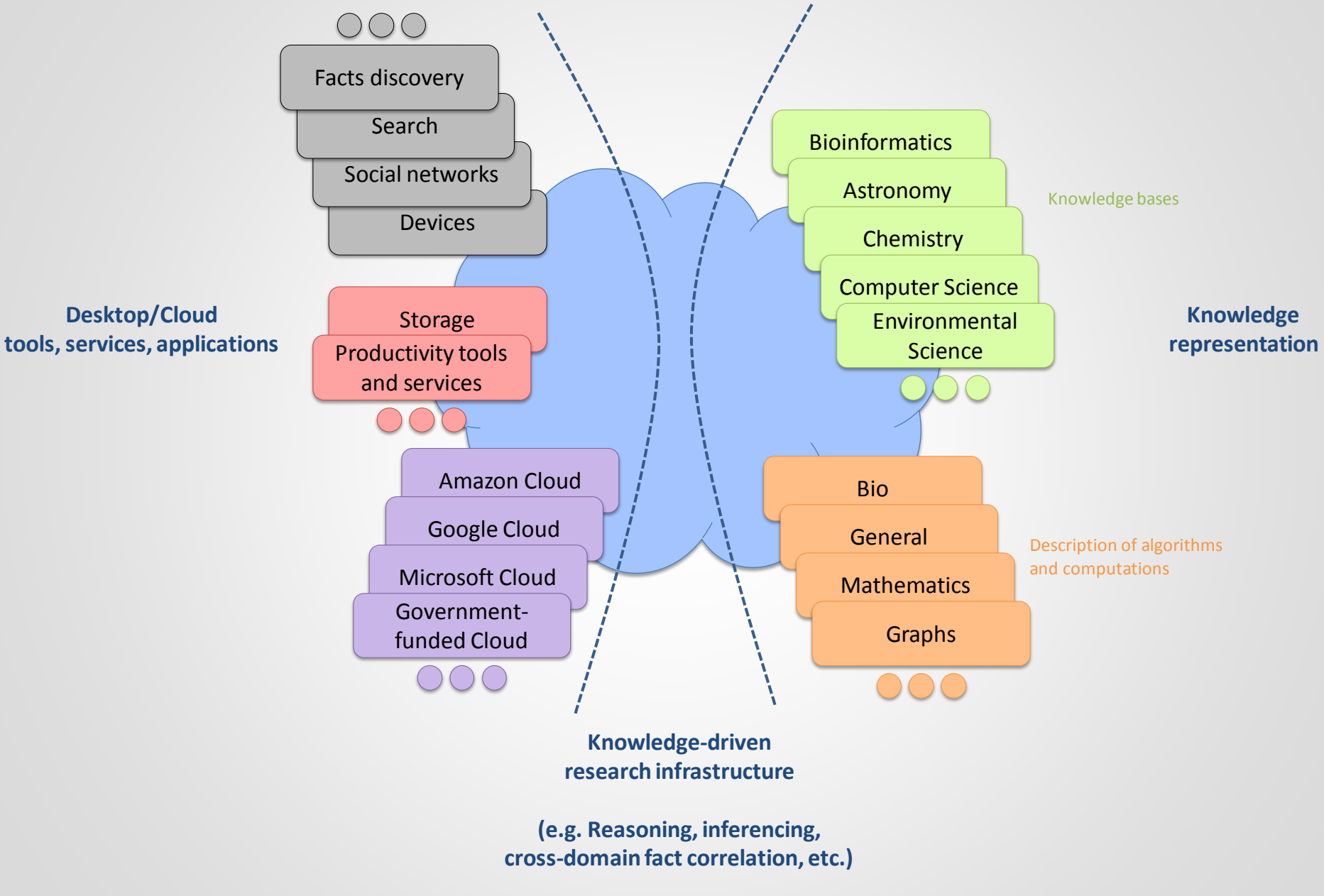
TOMORROW…

Computers are great **tools** in

huge amounts of **data**

We would like computers to also help with the **automatic**

of the world's **information**

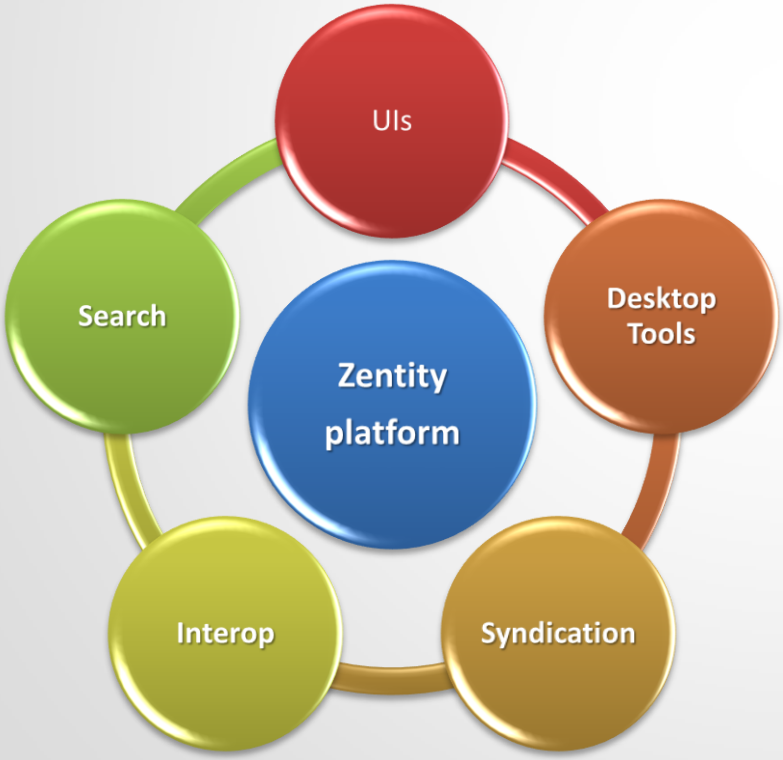

Facts discovery
Search
Social networks
Devices
Storage
Productivity tools and services
Amazon Cloud
Google Cloud
Microsoft Cloud
Government-funded Cloud
Bioinformatics
Astronomy
Chemistry
Computer Science
Environmental Science
Knowledge bases
Bio
General
Mathematics
Graphs
Description of algorithms and computations
Desktop/Cloud tools, services, applications
Knowledge representation
Knowledge-driven research infrastructure
(e.g. Reasoning, inferencing, cross-domain fact correlation, etc.)


Zentity

A platform for building services and tools for research output repositories

Papers, Videos, Presentations, Lectures, References, Data, Code, etc.

Relationships between stored entities



GOALS

Enable a tools and services ecosystem for “research output” repositories on MS technologies

Investigate concepts related to semantic computing

NON-GOALS

Support the lifecycle of publications

Compete with existing repository solutions

Modeling
RDFs – RDF Schema
Custom vocabularies (e.g. XML-based)
Direct use of extensibility API

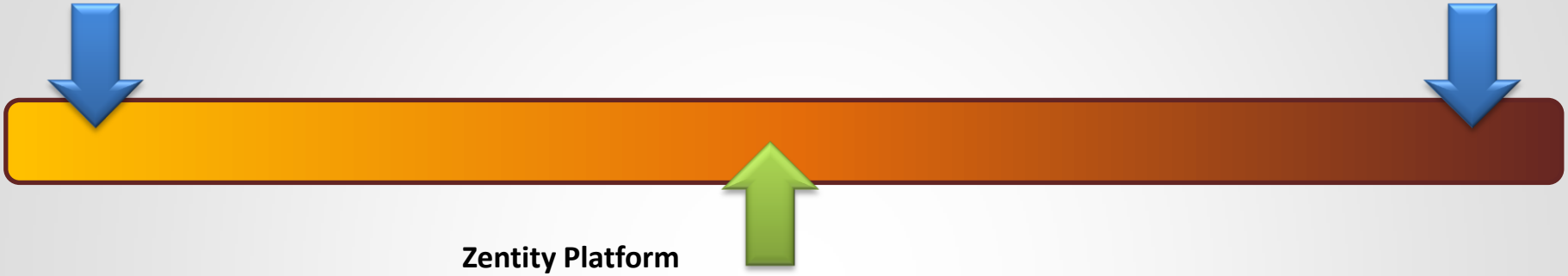
Syndication and Re-Use
RSS/Atom
OAI-PMH – Protocol for Metadata Harvesting
OAI-ORE – Object Re-Use and Exchange

Ingest & publishing protocols
SWORD – Simple Web-service Offering Repository Deposit
AtomPub
BibTex

HYBRID APPROACH

Triple stores
-Evolution friendly
-Poor performance
-No need to model everything in advance
-Semantic interpretation at the application level

Relational schema
-Evolution not so easy
-Great opportunities for optimization
-Model everything in advance



Zentity Platform
-Maintain a balance
-Try to model the frequently used entities in our app domain
-Try to capture the frequently used relationships
-Allow for extensibility (Relationships, Properties)

KEY DESIGN DECISIONS

Focus on “Resources”

Surface “Relationships” as first-class entities through our API

Model few relationships explicitly for performance reasons (e.g. “contains”, “author”, etc.) for our Scholarly Communications data model

Model key entities explicitly for our Scholarly Communications model

Expose the same functionality to arbitrary data models

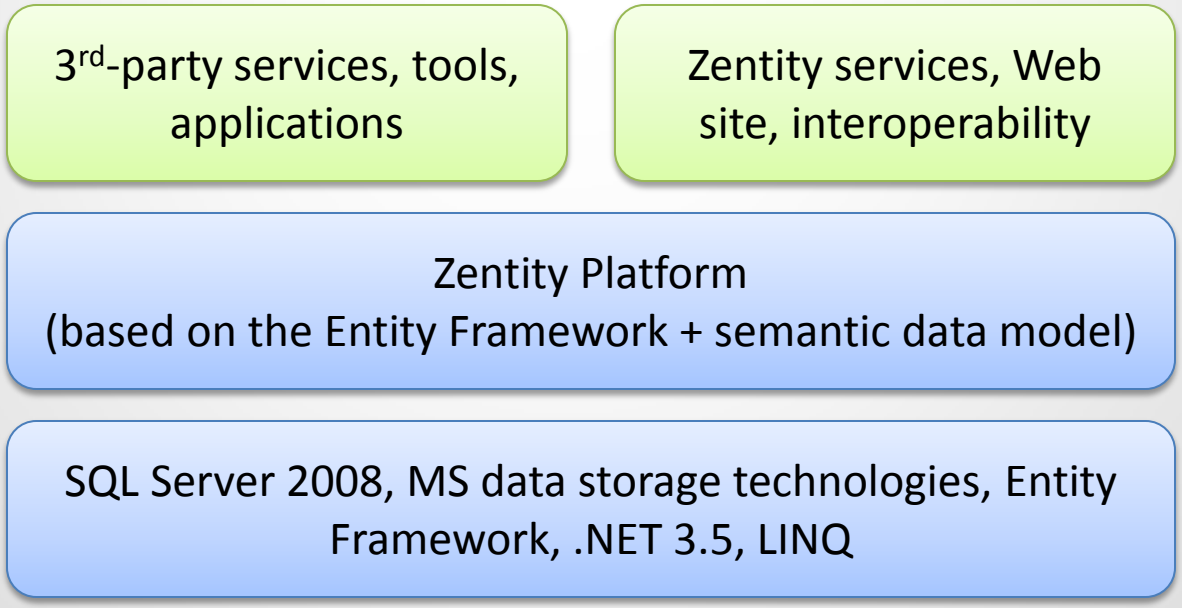
ARCHITECTURE GOALS

Create a platform for building “research output” repositories

An ecosystem of services and tools

Build an easy-to-install collection of basic services and tools

Extensibility



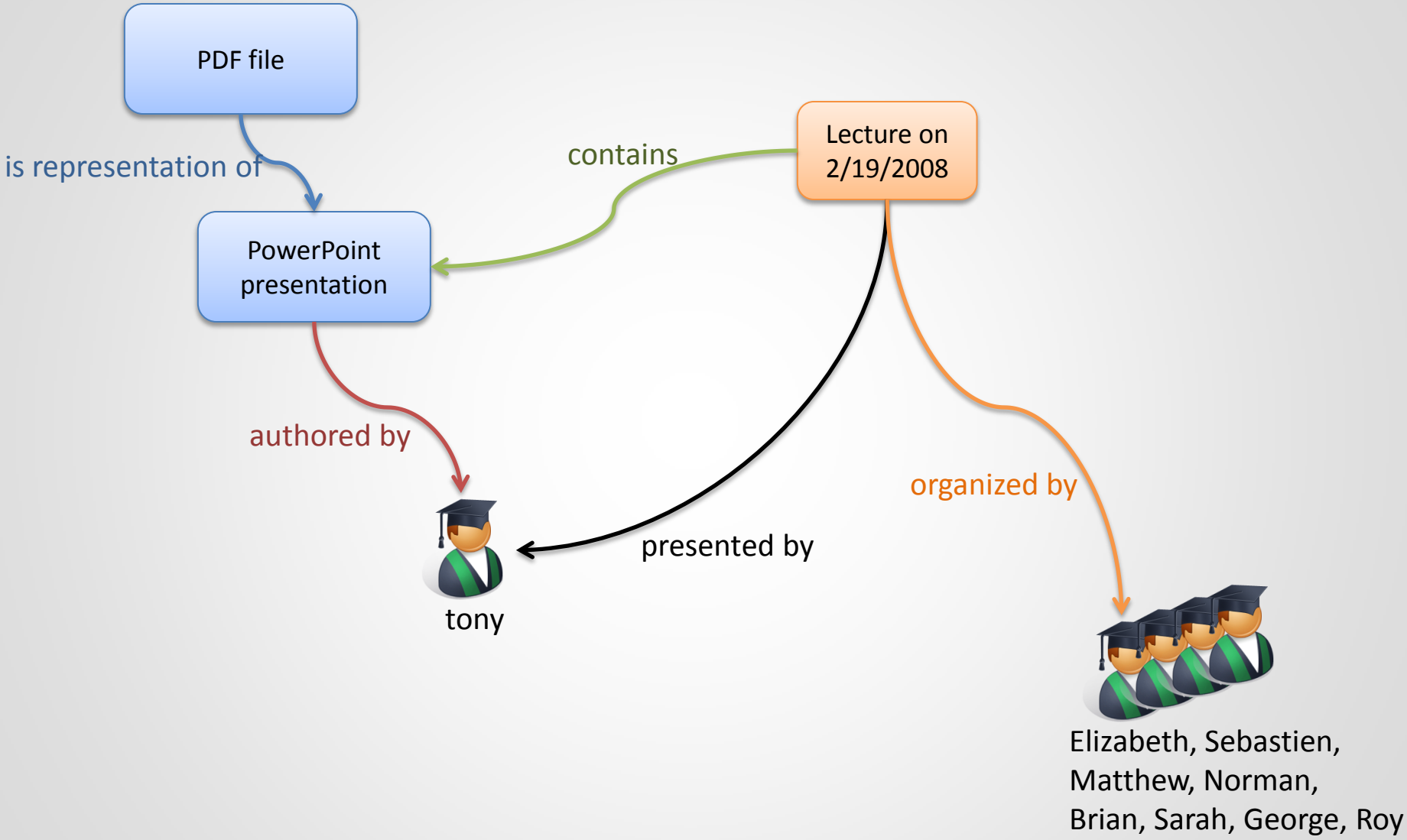
EXTENSIBILITY

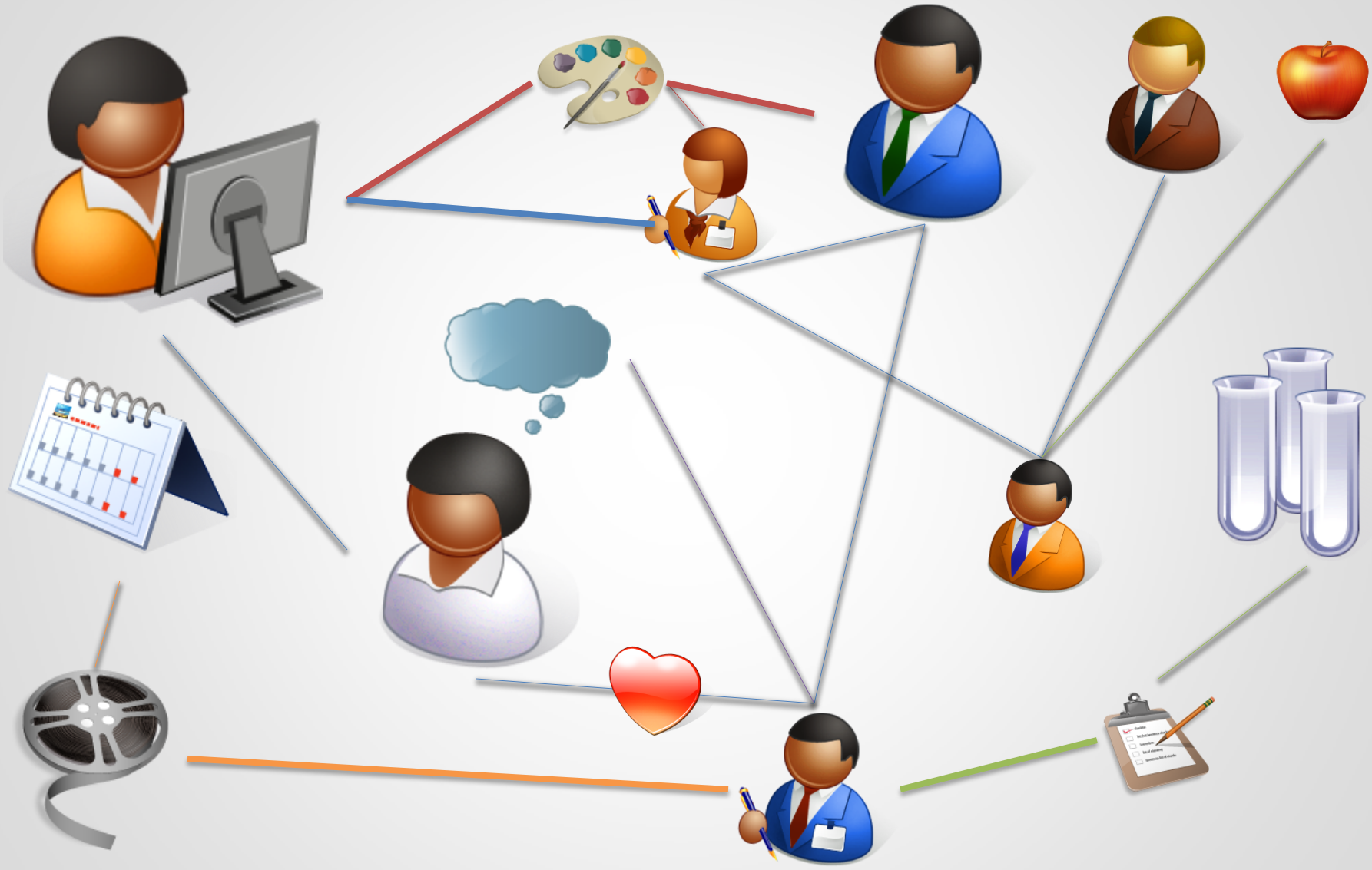
Declaratively described data models (e.g. DSL, RDFS)

Zentity Extensibility API

Store configured

.NET assembiles


PDF file
is representation of
PowerPoint presentation
contains
Lecture on 2/19/2008
authored by
presented by
organized by
tony
Elizabeth, Sebastien, Matthew, Norman, Brian, Sarah, George, Roy



Programming Experience

```
Person tony = new Person();

Publication pub1 = new Publication();
pub1.Title = "Title1";

Publication pub2 = new Publication();
pub2.Title = "Title2";

pub1.Cites.Add(pub2);
pub1.Authors.Add(tony);

Tag tag = new Tag();
tag.Name = "keyword";
pub1.Tags.Add(tag);
```

ROADMAP

CTP1 – Summer 2008

Beta 1 – Autumn 2008

RC1 – Winter 2008-2009

V1 – Spring 2009

Released in May 2009 at the Open Repositories conference!!!

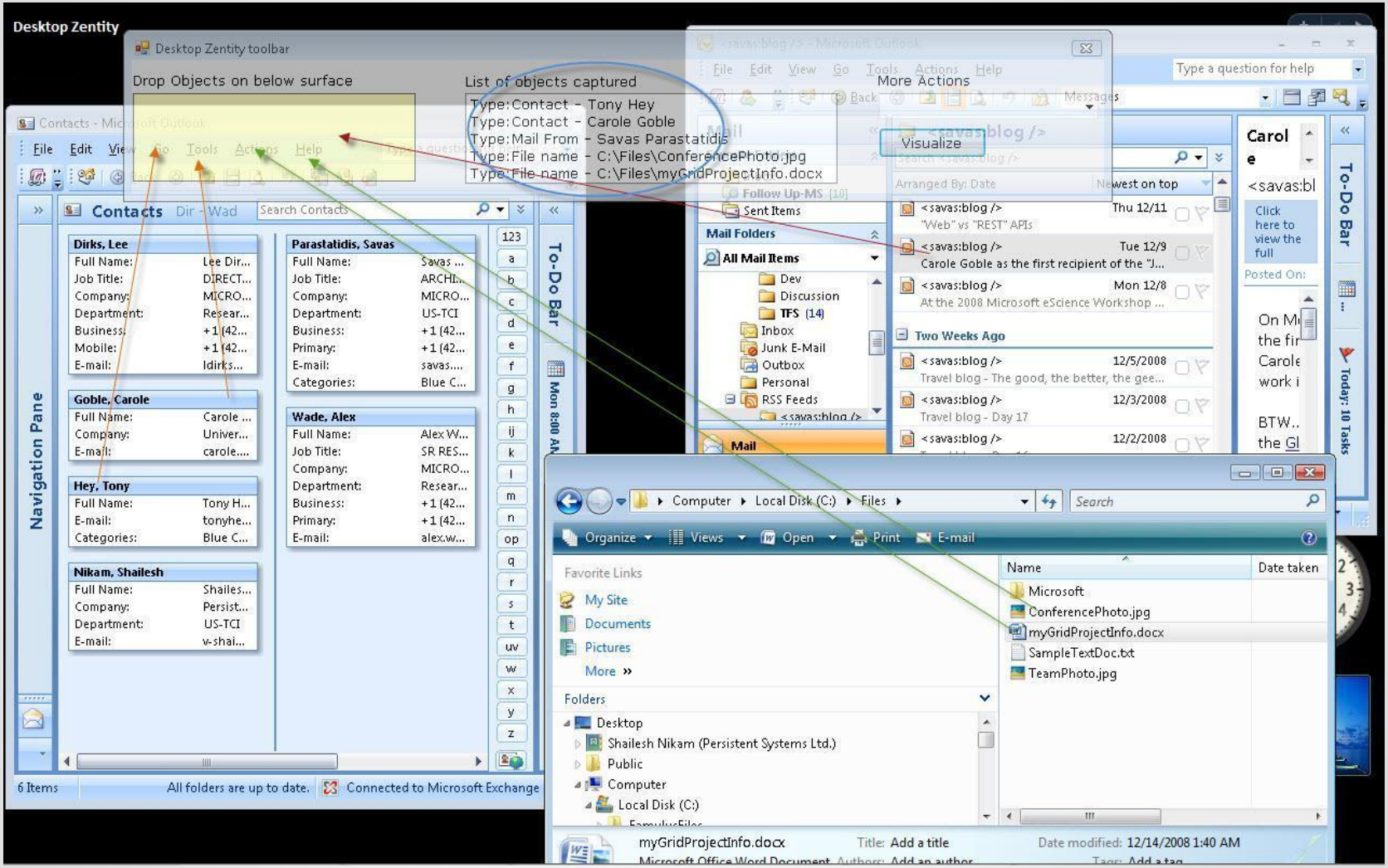
Next step: v1.1 and open source (provisionally: summer 2009)

DEMOS

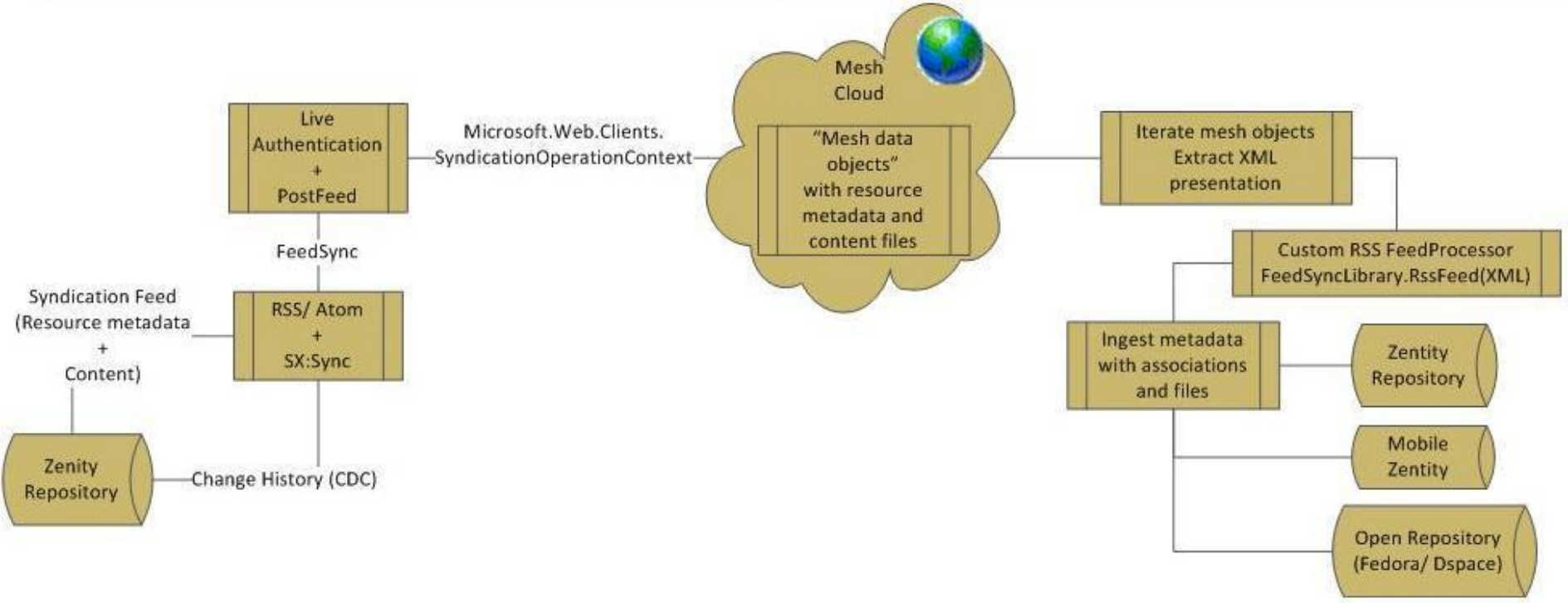
On a good day, 50% will NOT break :-)

Microsoft®

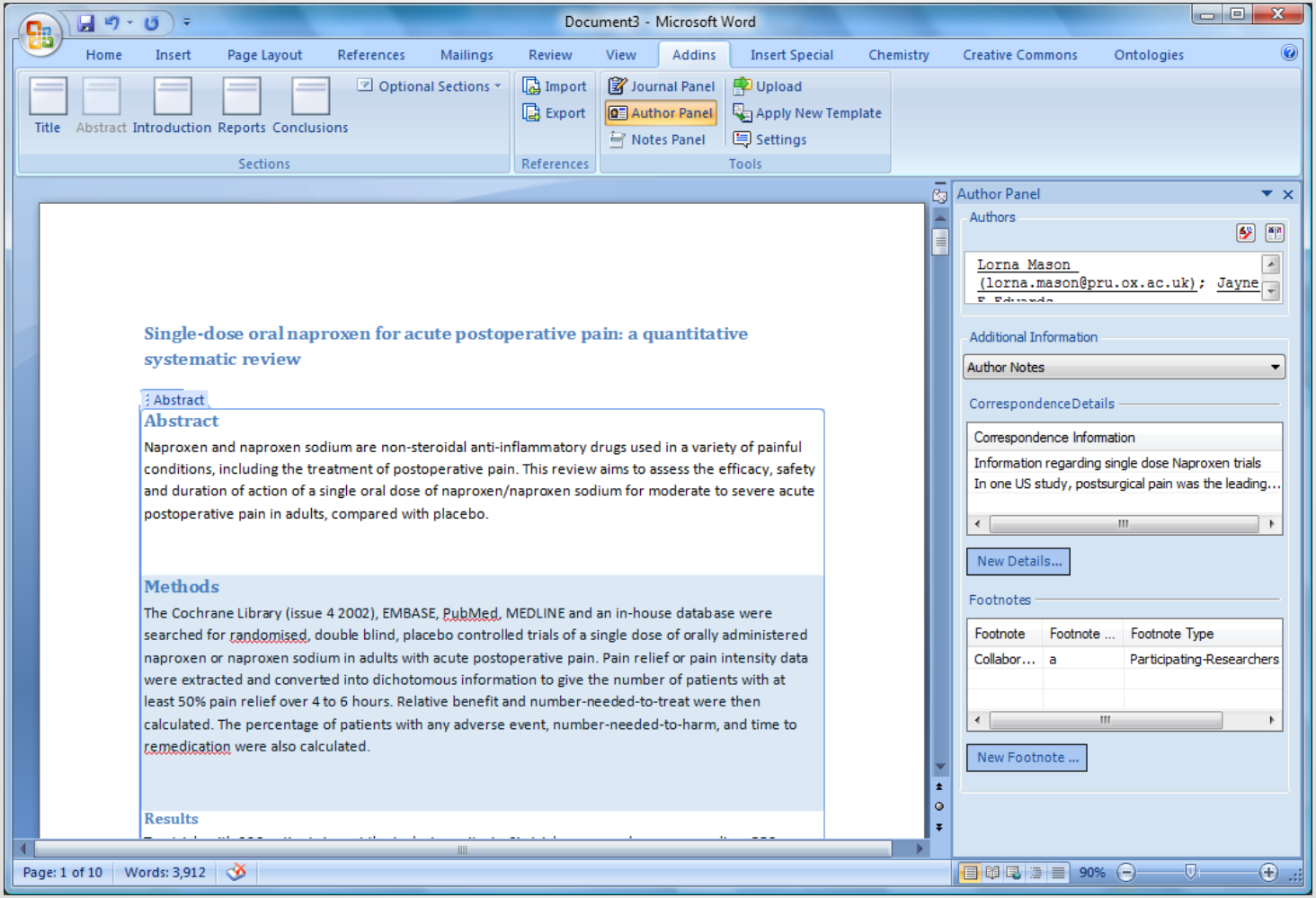
Your potential. Our passion.™


Desktop Zentity
Desktop Zentity toolbar
Drop Objects on below surface
List of objects captured
Type:Contact - Tony Hey
Type:Contact - Carole Goble
Type:Mail From - Savas Parastatidis
Type:File name - C:\Files\ConferencePhoto.jpg
Type:File name - C:\Files\myGridProjectInfo.docx
More Actions
Visualize
Contacts
Navigation Pane
Mail Folders
To-Do Bar
Computer
Local Disk (C:)
Files
Favorite Links
Folders
Name
Date taken
Microsoft
ConferencePhoto.jpg
myGridProjectInfo.docx
SampleTextDoc.txt
TeamPhoto.jpg
6 Items
All folders are up to date.
Connected to Microsoft Exchange


Zentity, Mesh & FeedSync
Lectures
Journals
Publications
Audio
Video
Presentations
Personal Zentity
Zentity Repository @ Organization
Mobile Zentity
Open Repository (Fedora/DSpace)


Live Authentication + PostFeed
Microsoft.Web.Clients. SyndicationOperationContext
Mesh Cloud
"Mesh data objects" with resource metadata and content files
Iterate mesh objects Extract XML presentation
FeedSync
Custom RSS FeedProcessor FeedSyncLibrary.RssFeed(XML)
Syndication Feed (Resource metadata + Content)
RSS/ Atom + SX:Sync
Ingest metadata with associations and files
Zentity Repository
Mobile Zentity
Open Repository (Fedora/ Dspace)
Zenity Repository
Change History (CDC)

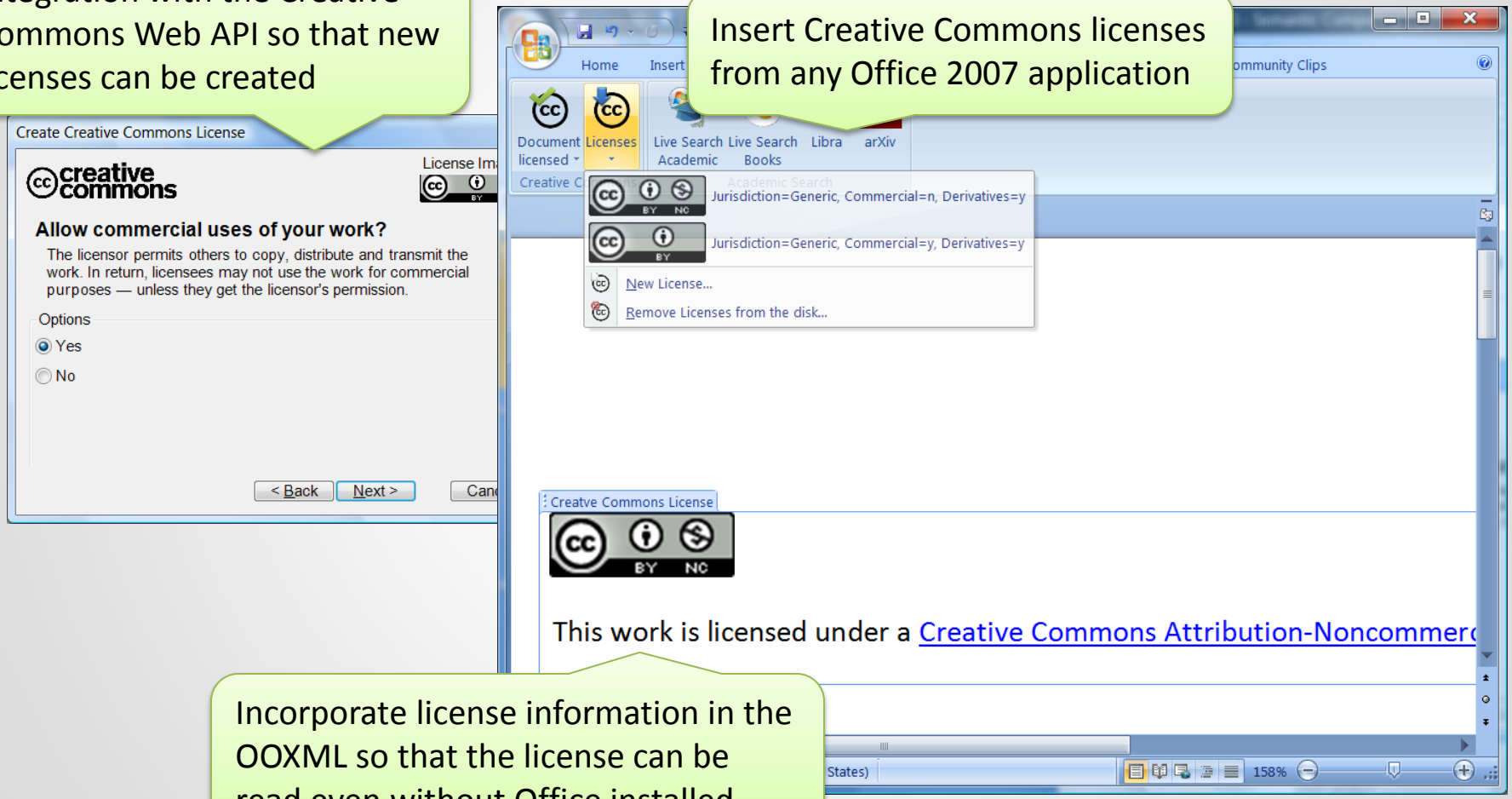
Article Authoring Add-in for Word 2007



CREATIVE COMMONS ADD-IN FOR OFFICE 2007

Integration with the Creative Commons Web API so that new licenses can be created

Insert Creative Commons licenses from any Office 2007 application



Incorporate license information in the OOXML so that the license can be read even without Office installed

Ontology Add-In for Word 2007

Execution

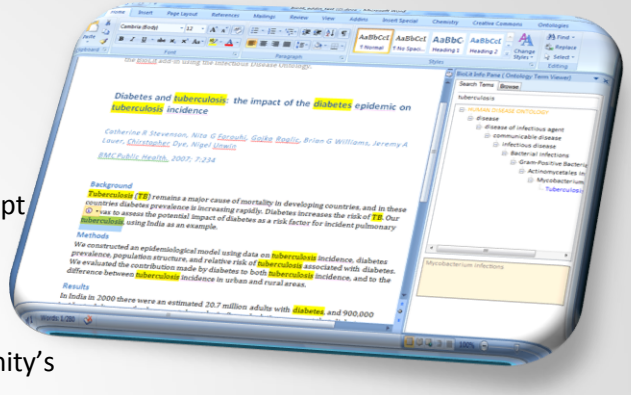
- Joint announcement between Microsoft Research & Creative Commons at O'Reilly eTech
- Binary and source code available on CodePlex as of 3/11/2009 (http://ucsdbiolit.codeplex.com/release/)
- Based on a research project with Dr. Phil Bourne at University of California-San Diego (2008)

Goals

- Facilitate semantic mark-up using ontologies and controlled vocabularies
- Facilitate/automate referencing to PDB, NCBO and other bio-related resources from manuscript

Scenario

- Authors do not need to be aware of the use of semantic technologies
- A domain-specific ontology is downloaded and made available from within Microsoft Word
- Authors can record their intention, the meaning of the terms they use based on their community's agreed vocabulary



Domain-specific ontology

Annotations travel with the document

Can be used to improve domain-specific discovery of information, cross-linking, etc.

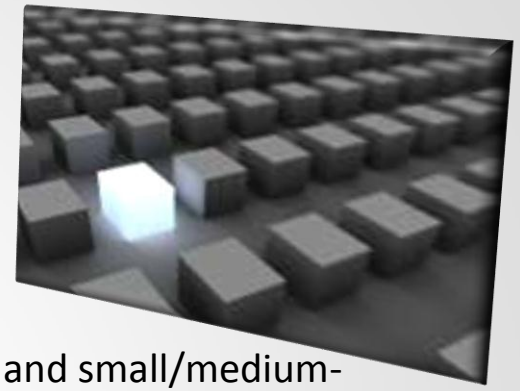
Support for annotations straight from within Word

creative commons

UCSD

eJournal Publishing Service

Hosted Offering for Scholarly Community



- With Microsoft Live Labs & Interoperability

Goals

- Provide a hosted full-service solution to support scholarly societies and small/medium-sized publishers–-simplify self-publishing of workshop/conference proceedings and small journals, as well as online collaboration between authors
- Make broader use of the peer review engine from MSR's Conference Management Tool (CMT) – assisting the review process
- Engage with the publishing ecosystem and enable third-party services (editorial, print-on-demand, etc.)
- Demonstrate interaction between CMT, eJournal Service, and Famulus (and other repositories)

Execution

- Initial focus on Science / Medicine
- Built on Word 2007, SharePoint, and Office Live
- Beta release in July 2008 – broad range of participants (10+)

Zentity

Research Output Repository Platform

A platform for building services and tools for research output repositories:

- Papers, Videos, Presentations, Lectures, References, Data, Code, etc.
- Relationships between stored entities

Goals

- Support the MSR publishing and dissemination platform for all researcher outputs
- Enable a tools and services ecosystem for “research output” repositories on MS technologies

Execution

- Utilizing OAI-ORE, SWORD, and other community protocols
- In development, deployment within MSR in early Q4
- Release to the community in late Q4
- Built on SQL Server 2008 + Entity Framework
    - Using WPF and Silverlight for UI