

Research-Output Repositories

An overview of Microsoft Initiatives



Our Commitment to Science

Putting computing into science...

Applying Microsoft products and research technologies to advance the scientific research and engineering innovation process

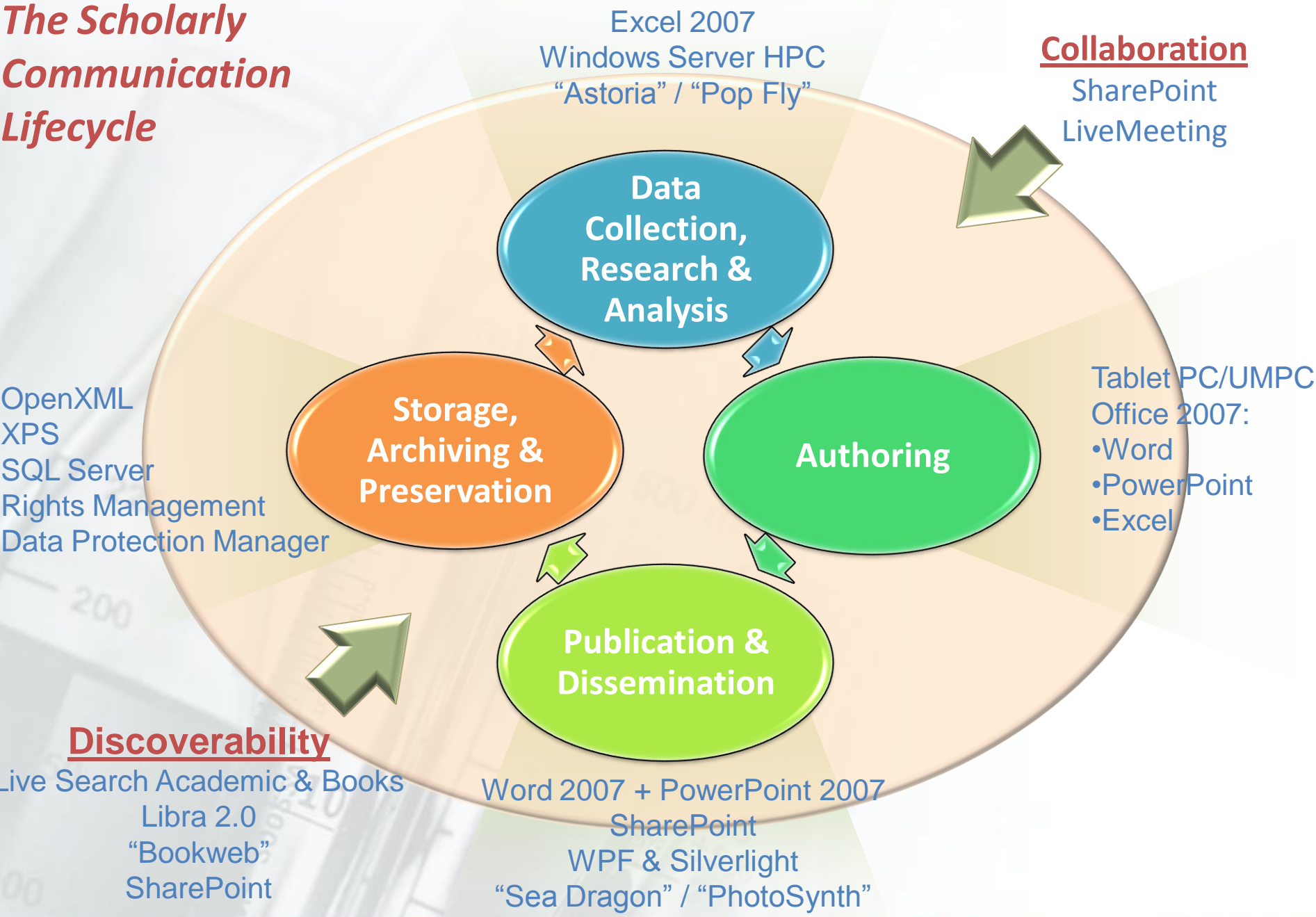
Putting science into computing...

Ensuring that research community requirements are factored into future versions of Microsoft software

- **Advancement of Science**
 - **Interoperability**
 - **Technology Excellence**
- **Global, Cross-Domain Collaboration**



The Scholarly Communication Lifecycle



Why Scholarly Communication?

- **Science + computation are not the entire equation**
 - Authoring, Analysis, Publishing, Discoverability, and Data Storage/Preservation are key components to scientists' everyday work...and Microsoft's core businesses
- ***The scholarly community has made it clear to us:***
 - Microsoft must improve its offerings throughout the scholarly communication lifecycle
- **MSR/TCI is uniquely positioned to drive this initiative within Microsoft**
- **Our approach:** Conduct prototyping projects and proofs-of-concept to evolve Microsoft's scholarly communication offerings



Audiences We Focus On

- **Academics / Scholars** (higher education setting)
- **Researchers / Scientists**
- **Libraries / Archives**
 - Academic, Research and National institutions
- **Scholarly Publishers & Societies**
 - Both Open Access and For-Profit enterprises
- **Governments / Related Organizations**
 - EU, NIH/NLM, NSF, NASA, etc.
 - JISC (UK), OCLC, CNI, DLF, NISO, etc.



Goal: *Transform Scholarly Communication*

- **Interoperability is paramount**
 - Actively lobby and drive for consensus around technical standards and standardized protocols proactively adopted by the community; enable broad community engagement
 - Customers have told Microsoft that the interoperability (and intellectual property) are OUR responsibility
- **Optimize for data-driven research & science** (open data/access)
 - To both data (scientific) and to information (scholarly publications)
 - Reproducible research + computational science
 - Properly document / annotate scholarly output
- **Data preservation (and provenance) should be baseline**
 - Documentation of the data's provenance
 - Reliable and secure long-term storage – at a massive scale
 - Preservation needs to be like “accessibility” features – i.e., assumed as required
- **Community protocols & conventions**
 - Leverage existing resources, tools, standards, or guidelines adopted by the community
- **Social networking & semantic knowledge discovery**
 - Harnessing collective intelligence must be a consideration – since accessing research is a core step in the life-cycle. Enable knowledge discovery
 - Optimize for Web 2.0 scenarios and allow end-users/experts to find things easier



Engagement Model: “Dual Benefit”

- **Work with researchers around the world**
 - Facilitate/advise on the application of technology
 - Link MSR researchers with (non-CS) researchers
- **Work with product groups**
 - Provide feedback on the use of MS technologies
 - Identify research-driven requirements for products
- **Terms & Conditions**
 - Microsoft typically shares IP (via BSD-type license) or makes source code available on <http://www.codeplex.com>
 - All projects result in freely available add-ins or downloads
 - Microsoft will not develop on a Linux platform
- **Project Execution Models**
 - a) Internal Development (FTE)
 - b) External Development (Vendor)
 - c) External Development (Institutional)
 - d) Mixed Model



Scholarly Communications: *Project Overview*

- **Current or Completed Projects**

- **Cornell** – arXiv.org + Word 2007 (and repository interoperability via SWORD)
- **MIT / Broad Institute** – Authoring (Word 2007) + data for research reproducibility
- **MSR** – CMT++ interoperability with data + metadata transfer/exchange (conference management tool enhancements)
- **LiveLabs** – eJournal publishing online service (community publishing tool)
- **UC San Diego / PLoS** – Semantic mark-up of scholarly articles (+ submission)
- **Chem4Word** with Office & Cambridge University – Create add-in to Word 2007 to facilitate drawing of chemical compounds and equations
- **Johns Hopkins University** – Digital Archive for Astronomy/Astrophysics data (storage, preservation and access)
- **Planets Project / EU** (with MSR – Cambridge) OpenXML and file format preservation + interoperability
- **eChemistry Project (Cornell, Penn State, Indiana, Cambridge, Southampton)** – ORE exemplar: access to compound chemical info objects (cross-repository access to open chemistry data)
- **British Library** – Researcher Information Centre (RIC) online workflow tool for scientists and researchers
- **Creative Commons Add-in for Office 2007** – evolving the Word 2003 effort
- **University of Southampton (UK)** – Port ePrints Repository Software for installation on the Windows platform
- **University of Manchester / “MyExperiment”** Project – social networking for scientists
- **ORE Acceleration Project** (OAI – Object Reuse & Exchange) – Alpha spec development
- **Indiana University** – Toolbox for Social Networking (SRT)
- **UK National Archives** – Virtual PC / Emulation of legacy systems to facilitate preservation
- **National Library of Medicine / NCBI** – “PubMed Int’l” UK version of PubMed + NLM DTD

- **Pipeline**

- **DRIVER 2 (EU)** – Infrastructure integration of across a network of European research repositories



The Scholarly Communication Lifecycle

Input / Linkage from TCI Science Projects

Collaboration

Researcher Info Centre
myExperiment

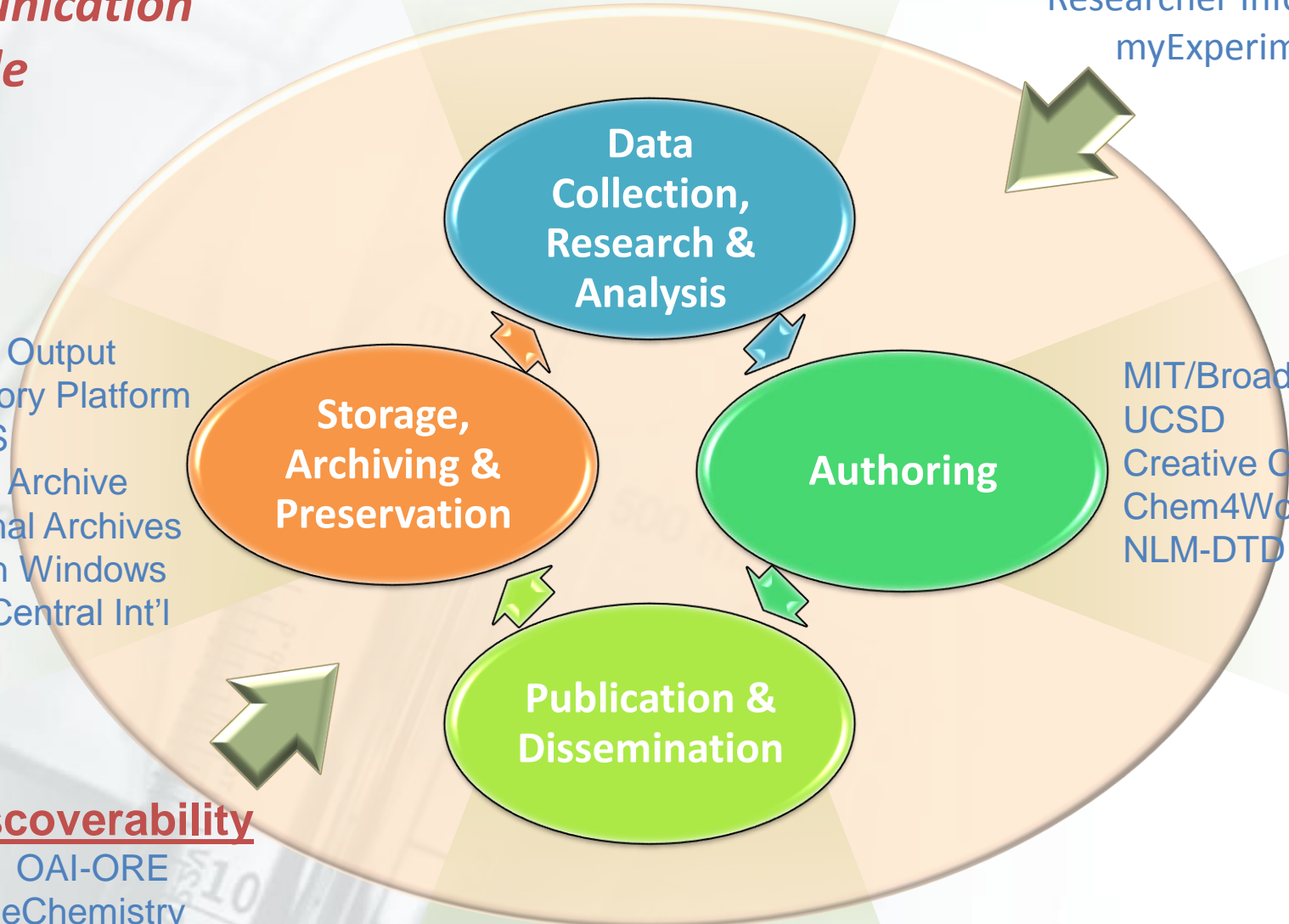
arXiv.org
Research Output Repository Platform
PLANETS
JHU Data Archive
UK National Archives
EPrints on Windows
PubMed Central Int'l

MIT/Broad
UCSD
Creative Commons
Chem4Word
NLM-DTD support

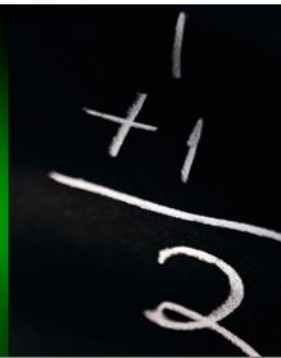
Discoverability

OAI-ORE
eChemistry
(Engagements with Libria 2.0 & "Bookweb" and planned work with FAST)

eJournal Hosted Publishing Service
CMT++



Microsoft Research Output Repository Platform



Research Output Repository Platform

Repository platform for storing scholarly works, and metadata about the scholarly works

- Papers, Videos, Presentations, Lectures, References, Data, Code, etc.
- Enables the development of new functionality and services on top of the platform
- Relationships between stored entities

Built on Microsoft Technologies
SQL Server, Entity Framework,
.Net Framework 3.5

Repository for Microsoft Research



Research Output Repository Platform

Goals

- Create a platform for building “research output” repositories
- Engage with the digital library and scholarly communications community
- Support an ecosystem of services and tools
- Available to the community for free (we are still considering the open source route)
- Build an easy-to-install collection of basic services and tools
- Inter-operability with existing systems, by implementing the community’s protocols and standards

Non-goals

- A generic platform for asset management
- Support the lifecycle of publications
- Compete with existing repository solutions



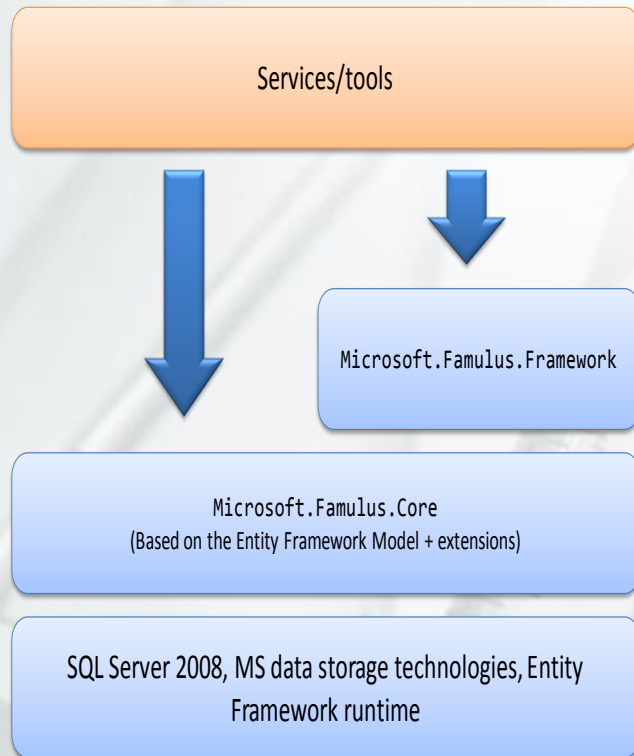
Research Output Repository Platform

Platform Overview

- Set of well-known resource types and associated properties that represents research output
 - e.g. Publications, technical reports, videos, presentations, data, lectures, files, person, organization etc
- Resource tagging
- Relationship between resources (Triple based)
- Set of well known predicates
 - e.g. IsVersionOf, Contains, IsRepresentationOf, AuthoredBy etc.
- New resource types and predicates through extensibility



Research Output Repository Platform



- Core API
- Framework API for ease of development
- Services
 - OAI-PMH, Syndication, BibTeX import/export, Search
 - UI Web Controls

Research Output Repository Platform

- A Semantic Computing platform
- A hybrid between a relational database and a triple store

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



Research Output Repository 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, Attributes)



An intuitive programming experience

```
Person tony = new Person();

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

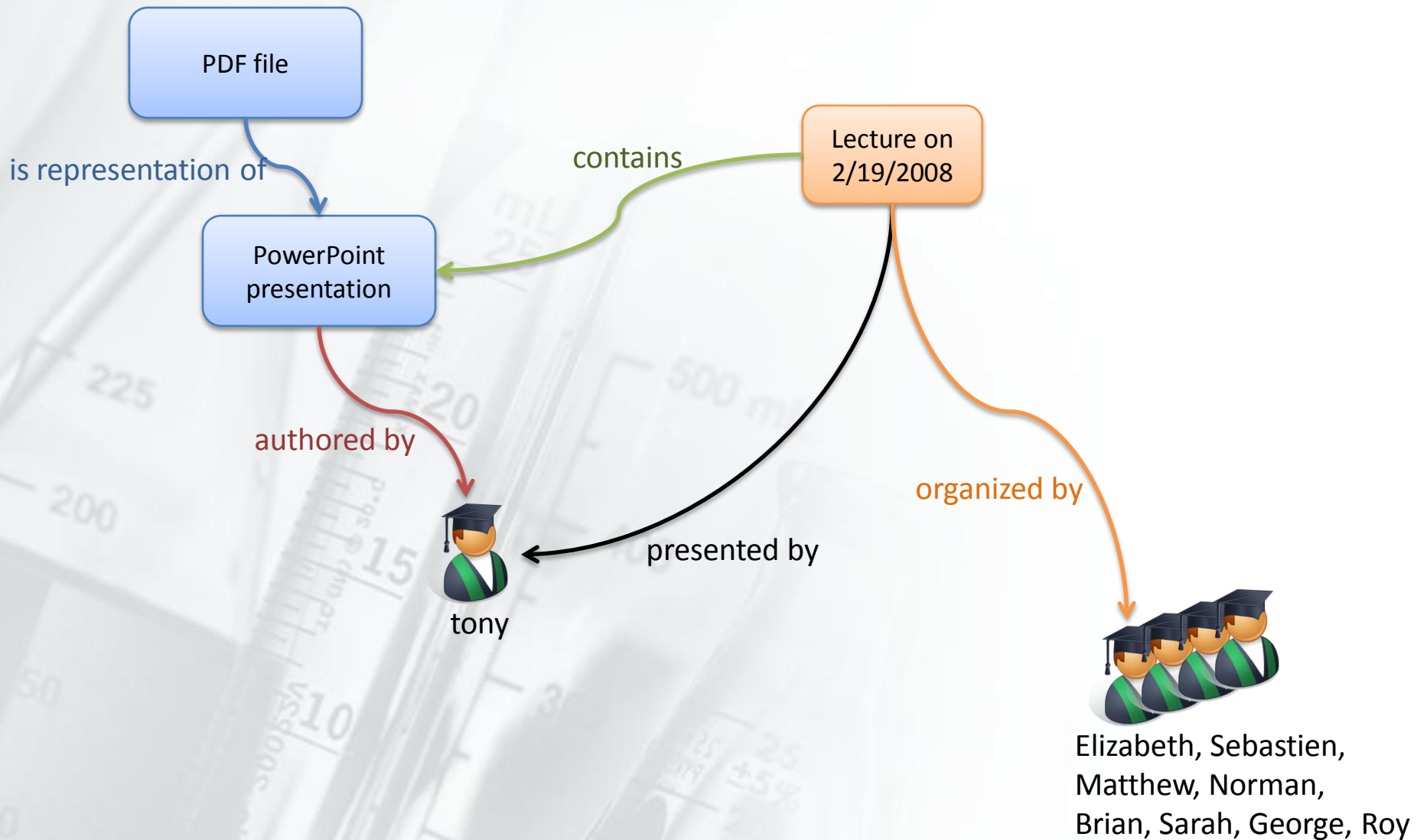
Publication pub2 = new Publication();
pub2.Title = "Title2";

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

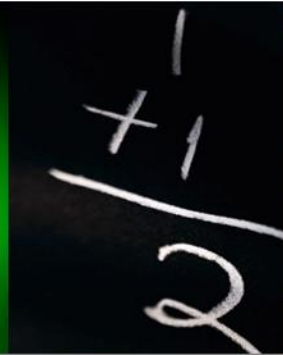
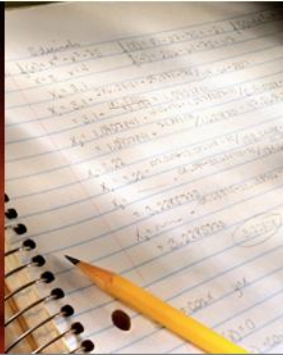
Tag tag = new Tag();
tag.Name = "keyword";
pub1.Tags.Add(tag);
```



Research Output Repository Platform



Demo



Research Output Repository Platform

Where we currently are:

- Nearing end of M1 phase of the platform
Expected completion by mid-April
- Ready for consumption by Microsoft Research platform
- Planning for M2 phase of the project

Probable M2 features:

- Change History
- Notifications and Logging
- Workflow Engine
- Additional Search features
- Web Service Interface
- Additional UI controls
- Aiming for a public beta release



Research Output Repository Platform

Forum for feedback, questions and discussions:

<http://community.research.microsoft.com/forums/90.aspx>



Contacts

- Lee Dirks, Director of Scholarly Communication (ldirks@microsoft.com)
- Alex Wade, Program Manager (awade@microsoft.com)
- Santosh Balasubramanian, Program Manager (santoshb@microsoft.com)
- Savas Parastatidis, Architect (savasp@microsoft.com)





Microsoft[®]

Your potential. Our passion.[™]

