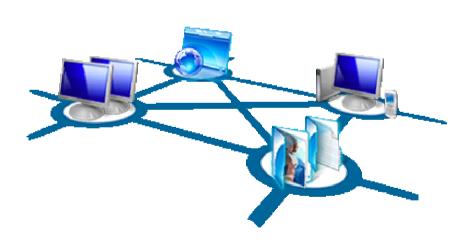
Tools for Dataset Lifecycle Management



Alex D. Wade Dean Guo Simon Mercer Oscar Naim Mike Zyskowski

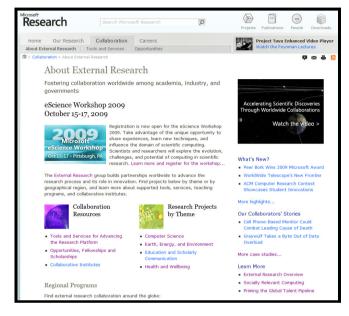
Research

Microsoft 1325211011



Microsoft External Research

- Division within Microsoft Research focused on partnerships between academia, industry and government to advance research in fields that rely heavily upon advanced computing
- Supporting groundbreaking research to help advance human potential and the wellbeing of our planet
- Developing advanced technologies and services to support every stage of the research process
- Microsoft External Research is committed to interoperability and to providing open access, open tools, and open technology





http://research.microsoft.com/collaboration/about/



Mission

Tailor Microsoft software to meet the specific needs of the academic research community

Our approach:

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



The

FOURTH PARADIGM

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

Free PDF Download Or, Amazon Kindle version & paperback print-on-demand

"The impact of Jim Gray's thinking is continuing to get people to think in a new way about how data and software are redefining what it means to do science."

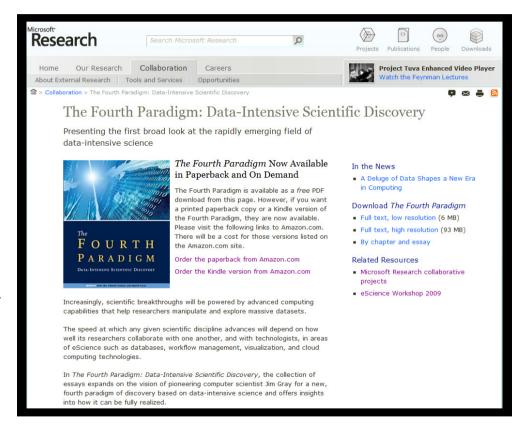
— **Bill Gates,** Chairman, Microsoft Corporation

"One of the greatest challenges for 21st-century science is how we respond to this new era of data-intensive science. This is recognized as a new paradigm beyond experimental and theoretical research and computer simulations of natural phenomena—one that requires new tools, techniques, and ways of working."

Douglas Kell, University of Manchester

"The contributing authors in this volume have done an extraordinary job of helping to refine an understanding of this new paradigm from a variety of disciplinary perspectives."

- Gordon Bell, Microsoft Research



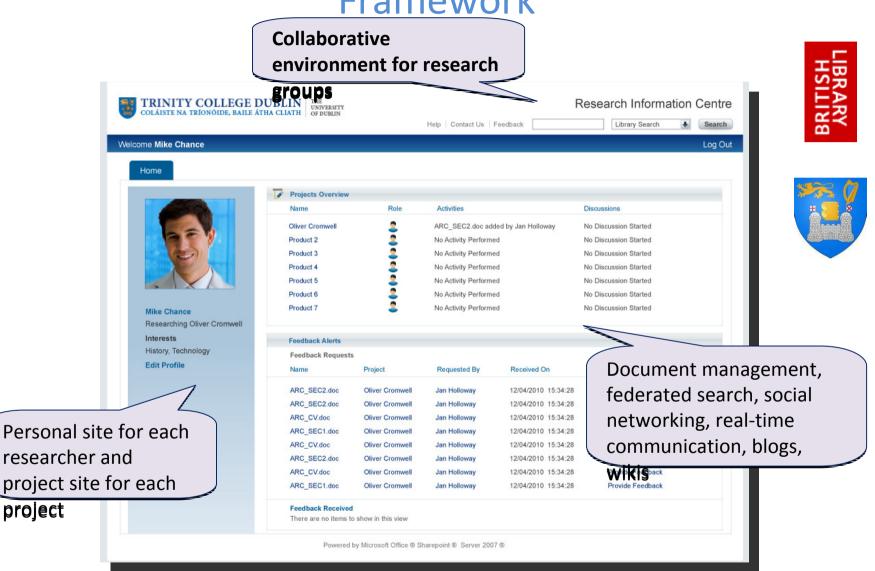


http://research.microsoft.com/fourthparadigm/

RESEARCHER PRODUCTIVITY

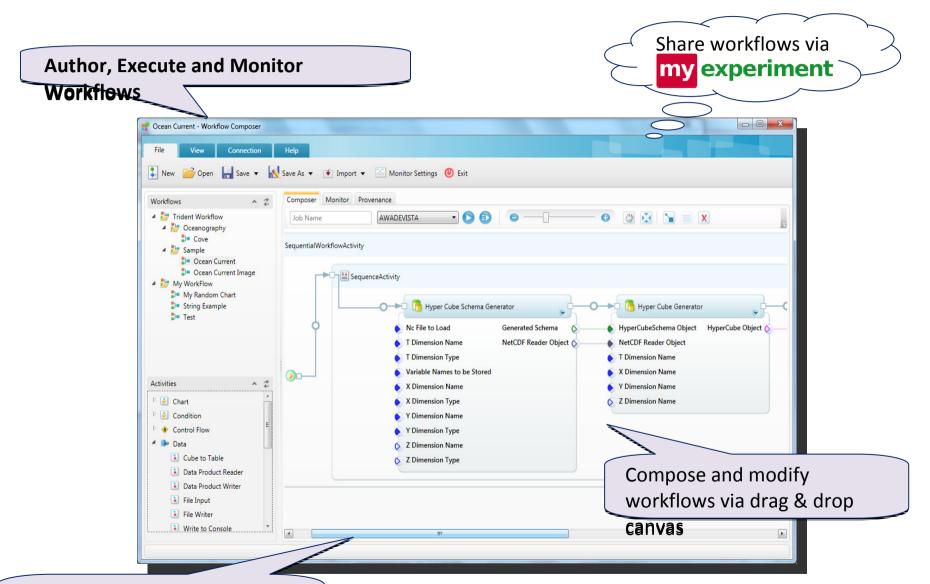
Make complex tasks simple

Research Information Centre – a VRE Framework



Version 1.0 (Open Source under Ms-PL): http://ric.codeplex.com/

Project Trident – a Scientific Workflow Workbench

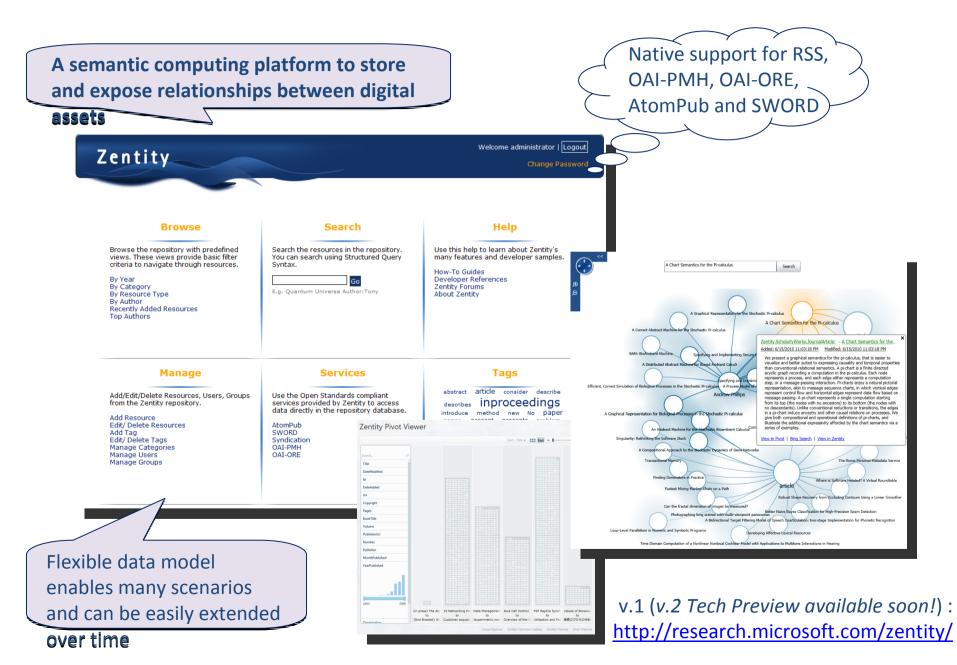


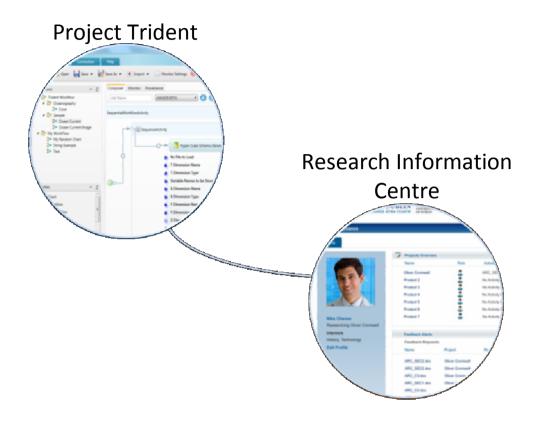
View data products, performance metrics, and

arailananea data

Version 1.2 (Open Source under Apache 2.0 License): http://tridentworkflow.codeplex.com/

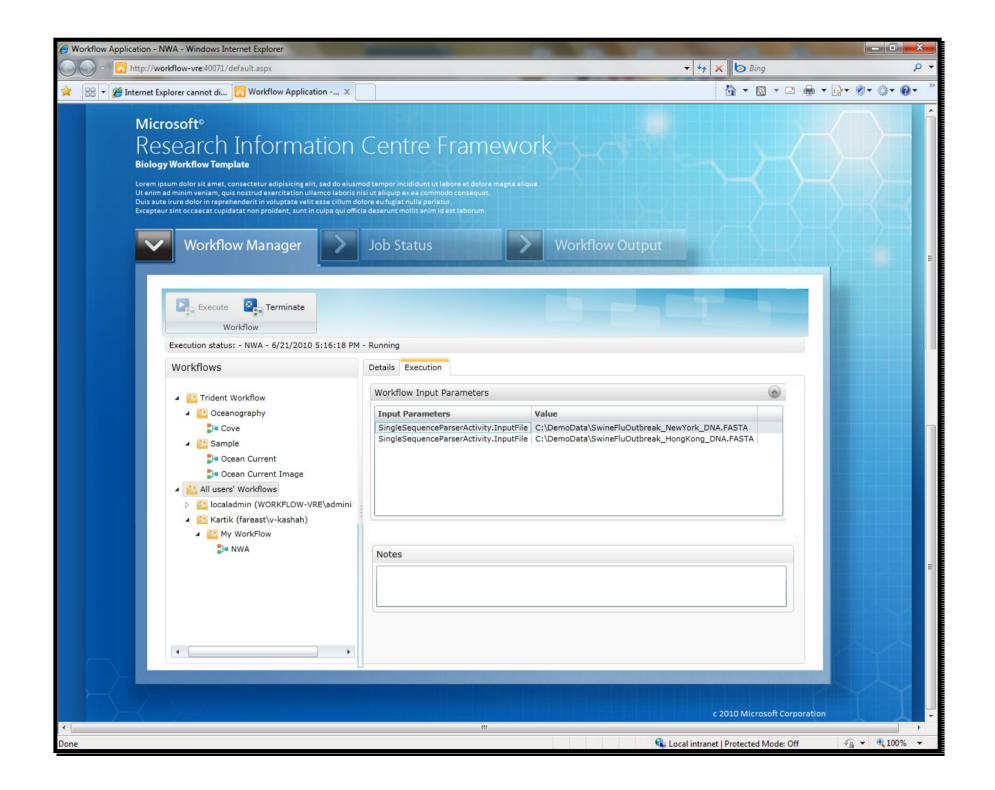
Zentity – a Research Output Repository Platform

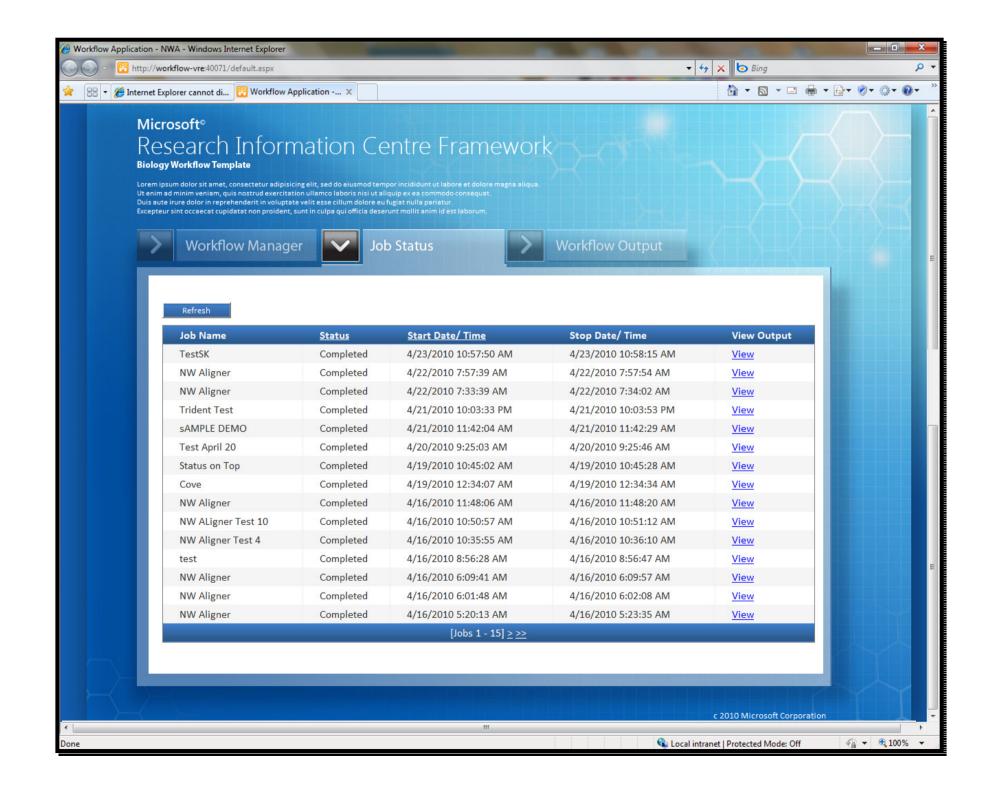


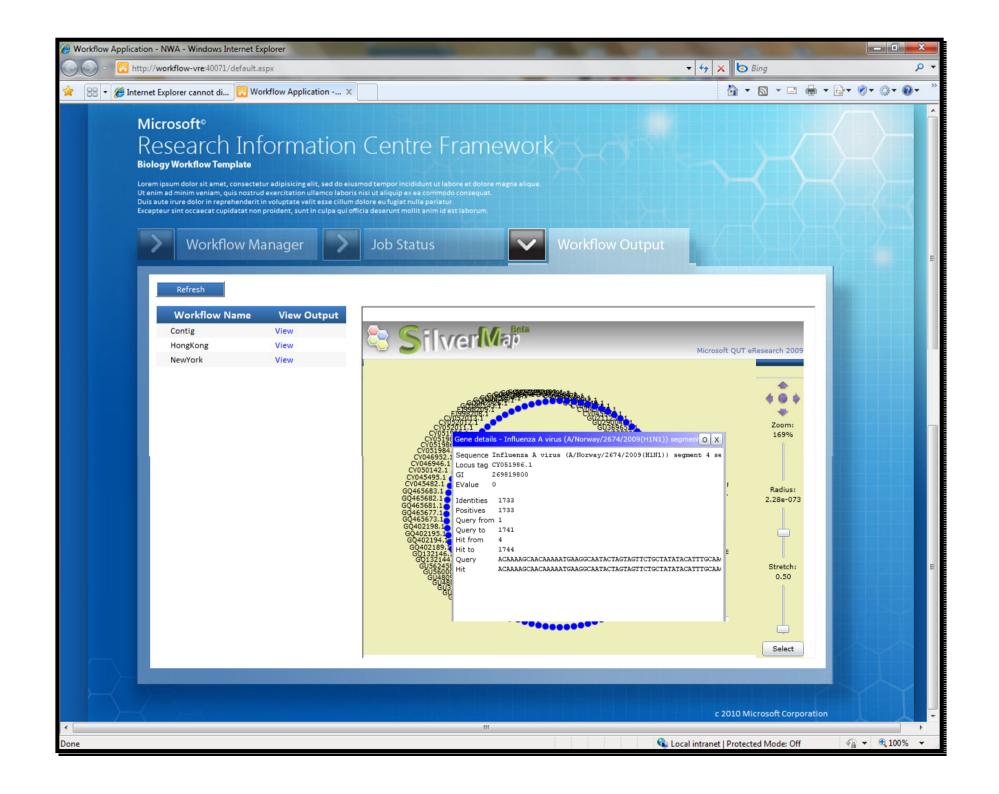


DATA MANAGEMENT

Integrated Workflow Template for the RIC





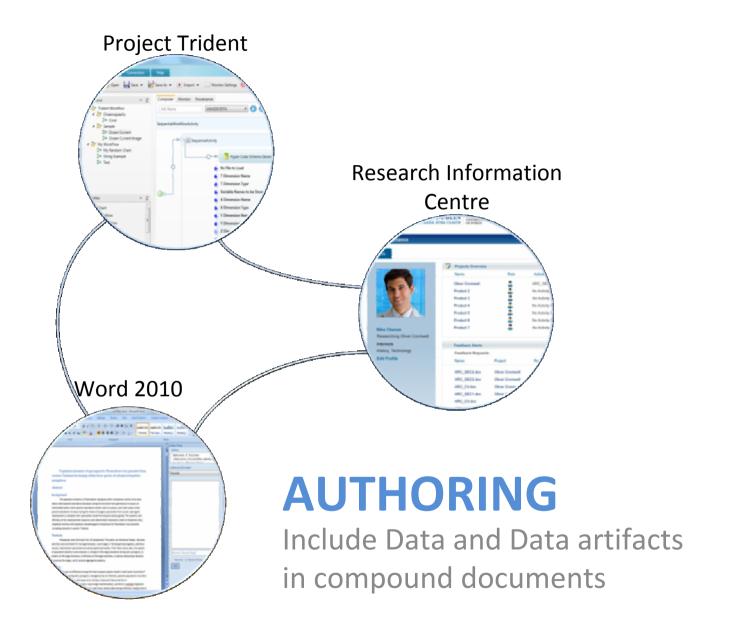


Integrated Workflow Template

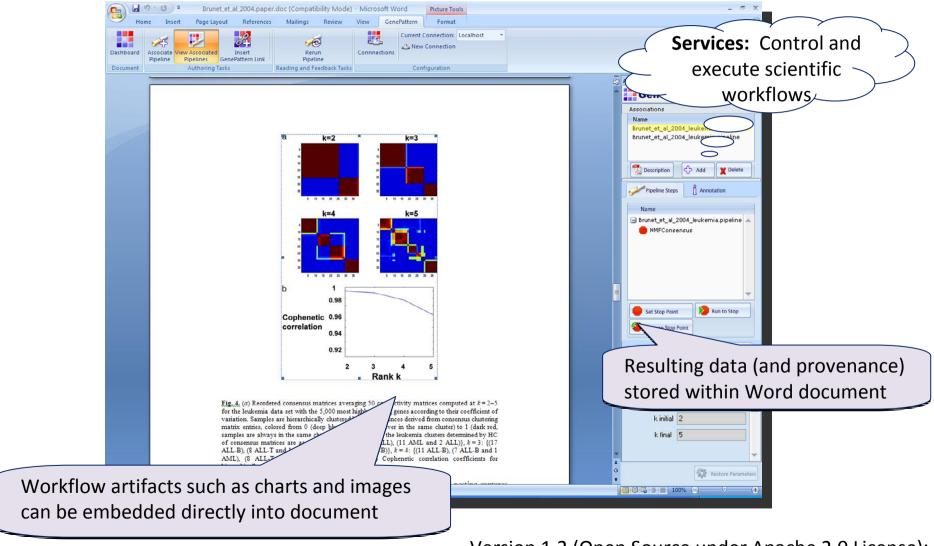
for the Research Information Centre

From the RIC, researchers can...

- View/execute/monitor scientific workflows within the context of project collaboration site
- Receive alerts (email, SMS) when workflows complete
- Browse workflow execution history and provenance information
- Review/store/manage data files that are written back into SharePoint by Trident

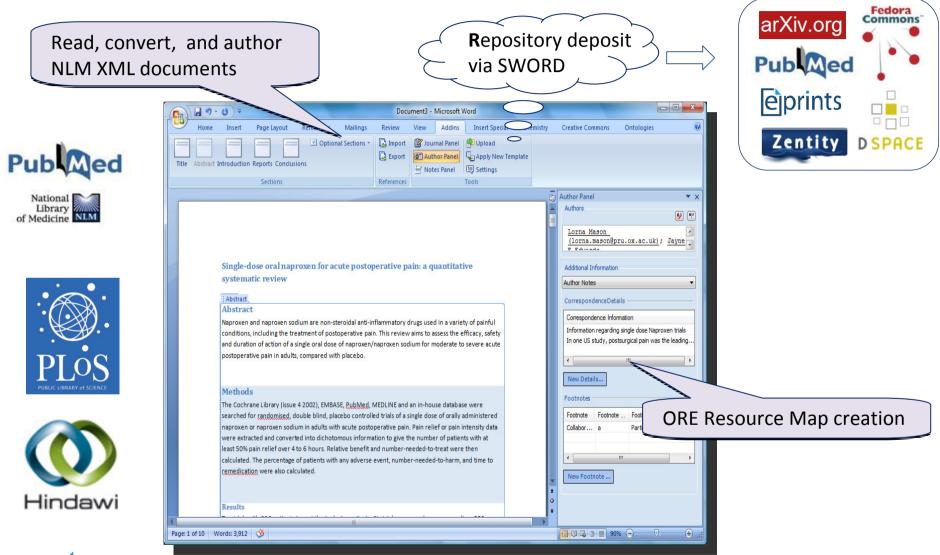


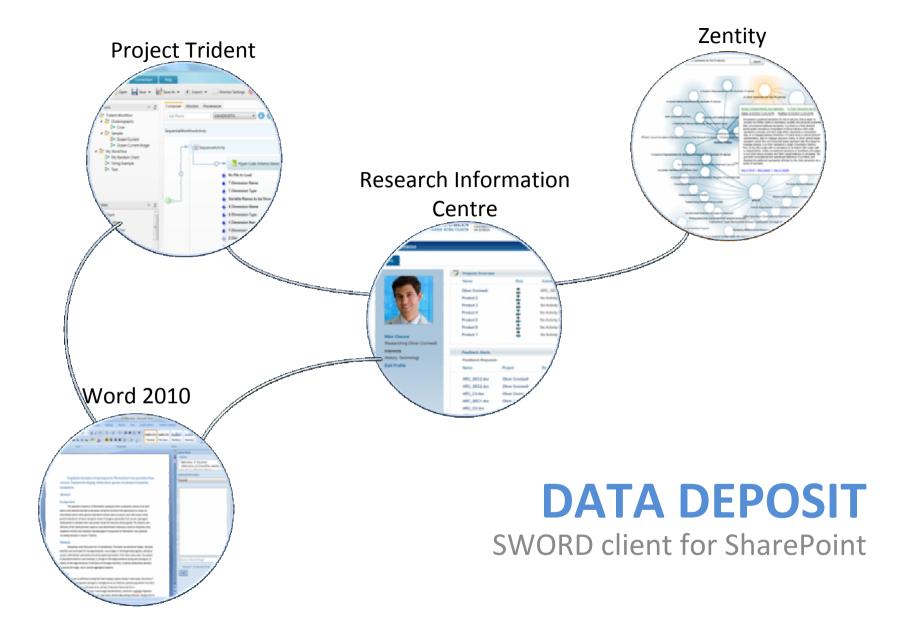
Project Trident: Workflow Add-in for Word



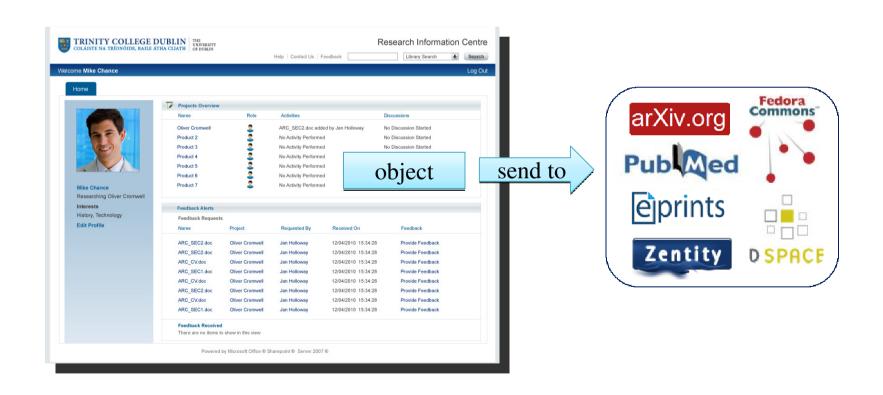
Version 1.2 (Open Source under Apache 2.0 License): http://tridentworkflow.codeplex.com/

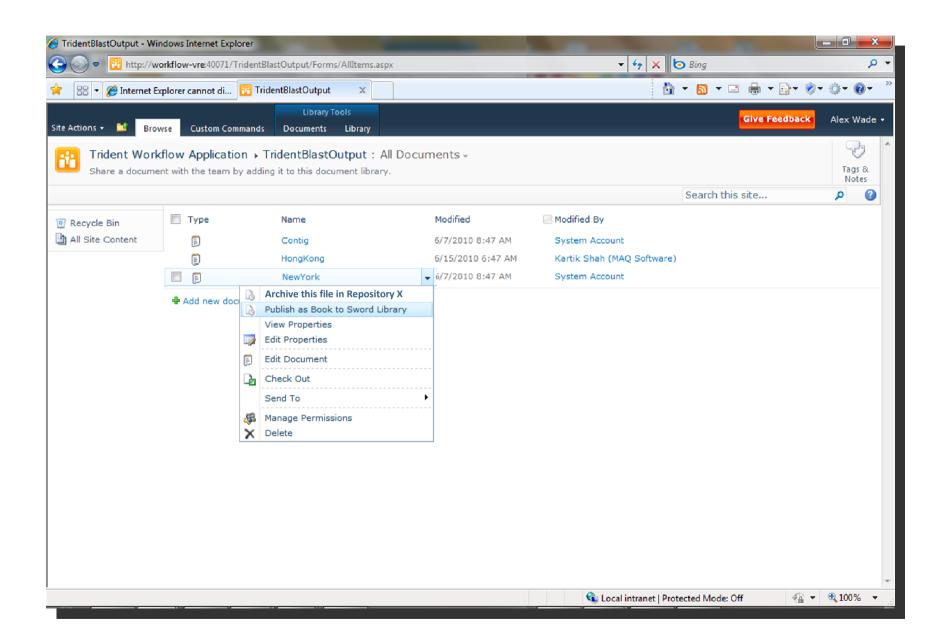
Article Authoring Add-in for Word





SWORD client for SharePoint



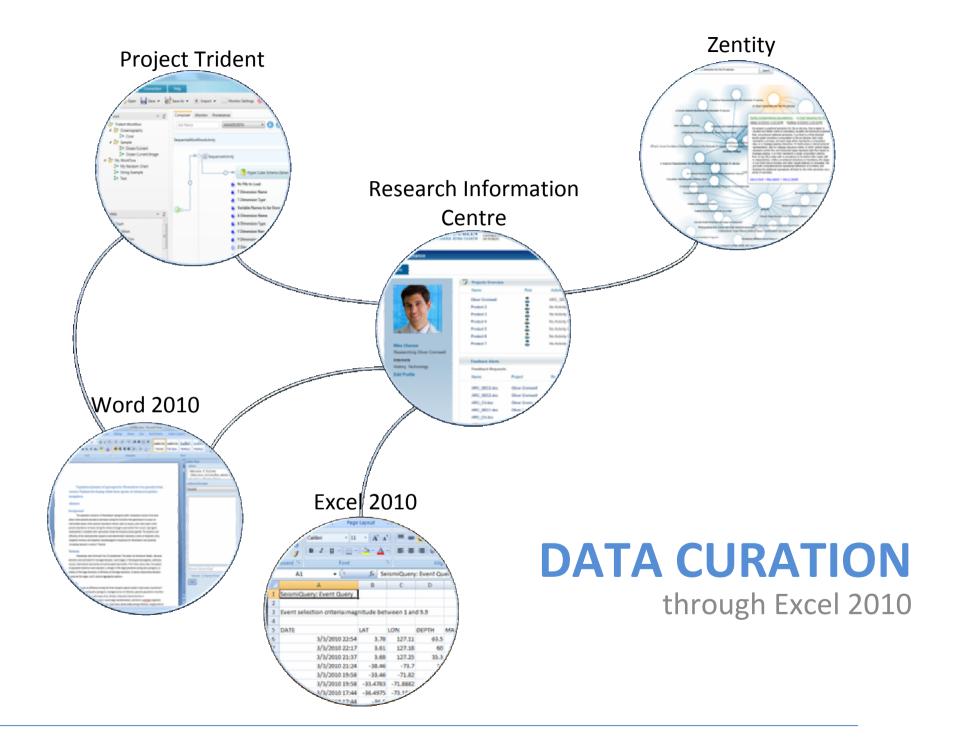


SWORD client for SharePoint



From SharePoint (and the RIC), researchers can...

- Select any file in SharePoint:
 - Document
 - Presentation
 - Image
 - Data files
- and publish it to any repository (via SWORD)
 - SWORD endpoints are managed as a custom list, so new locations are easily added



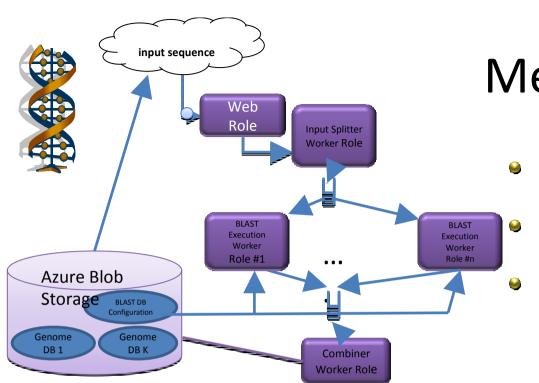
Data Curation Add-in for Microsoft Excel







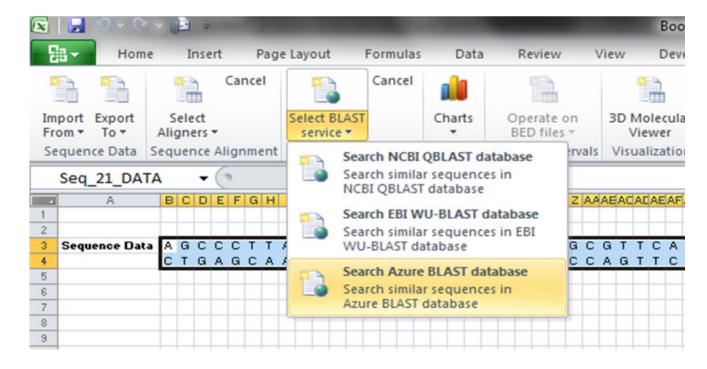
- Microsoft Research, in partnership with <u>California Digital Library's Curation Center</u>
 - Collaboration with Tricia Cruse & John Kunze
 - Part of the DataONE (an NSF DataNet Project)
- Proposed functionality under consideration:
 - Versioning revision history and original raw data can be protected and recovered
 - Time stamps easily determine when the data were created and last updated
 - "Workbook builder" select from globally shared standardized layouts for capturing data
 - Export metadata in a standard formats (e.g., a DataCite citation or an EML document that describes the
 dataset(s) in a workbook) so that researchers can readily share their data,
 - Globally shared vocabulary of terms for data descriptions (e.g., column names), and as needed to add new terms to the globally shared vocabulary, to enable wide collaboration between researchers
 - Import term descriptions from the shared vocabulary and annotate them to refine local definitions
 - Deposit data and metadata into a data archive to preserve and publish research data

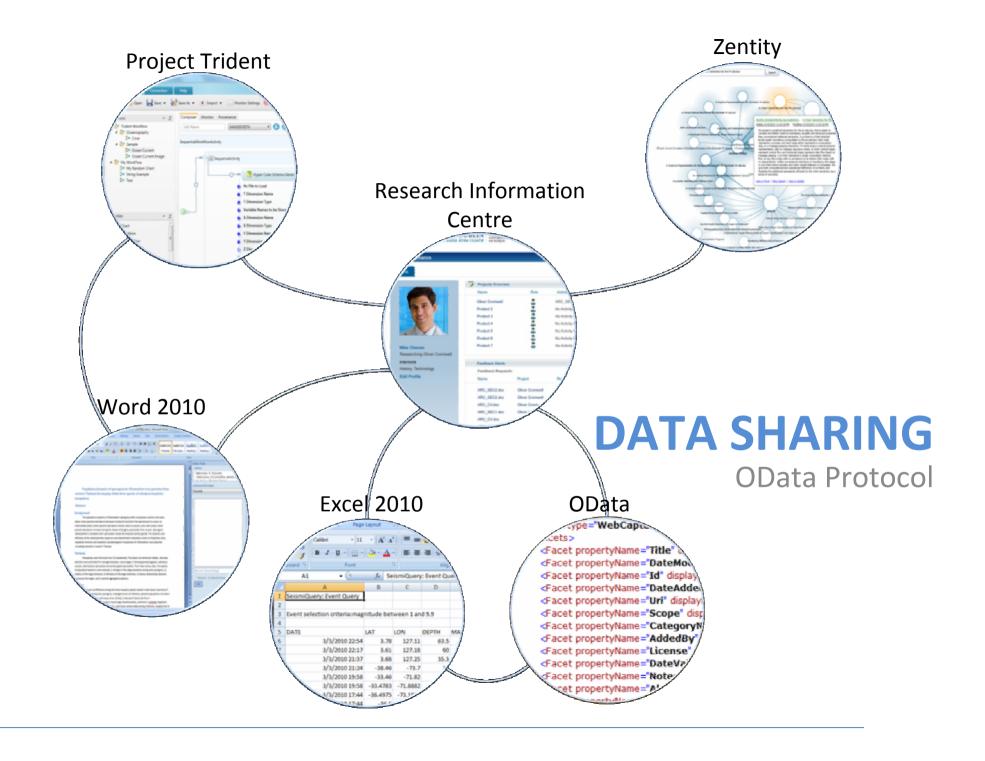


Metagenomics

- Using Azure cloud to host standard sequence data
- Run BLAST in parallel "map reduce" style for a large collection input
- Use Excel-based tool to search databases

- Performance: speedup
- 45x with 50 roles; 94x with 100 roles;
- with 300 roles computation time is reduced to 4 hours!







Just HTTP

- Items as resources, HTTP methods (GET, PUT, ...) to act
- Leverage proxies, authentication, ETags, ...

Uniform URL convention

- Every piece of information is addressable
- Predictable and flexible URL syntax

Multiple representations

- Use regular HTTP content-type negotiation
- JSON and Atom (full AtomPub support)

Open Data Protocol URL Conventions

Addressing lists and items

List of lists	/_vti_bin/listdata.svc/
List	listdata.svc/Employees
Item	listdata.svc/Employees(123)
Single column	listdata.svc/Employees(123)/Fullname
Lookup traversal	listdata.svc/Employees(123)/Project
Raw value access	listdata.svc/Employees(123)/Project/Title/\$value

Presentation options

Sorting	listdata.svc/Employees?\$orderby=Fullname
Filtering	listdata.svc/Employees?\$filter=JobTitleeq 'SDE'
Projection	listdata.svc/Employees?\$select=Fullname,JobTitle
Paging	listdata.svc/Employees?\$top=10&\$skip=30
Inline expansion	listdata.svc/Employees?\$expand=Project

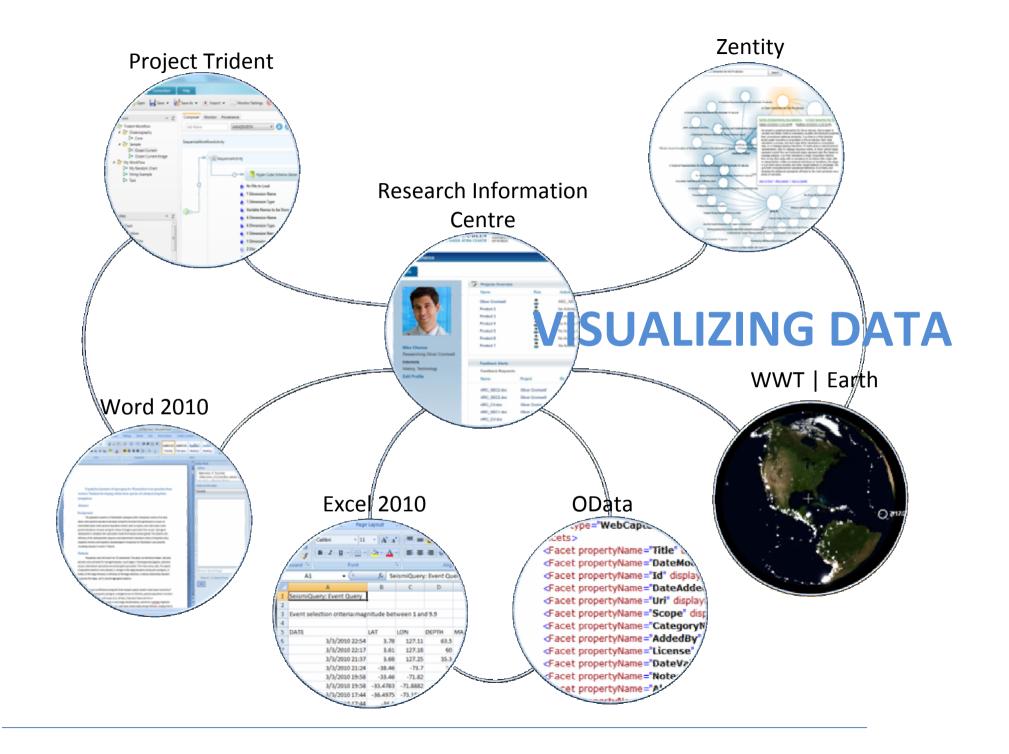


OData Producers

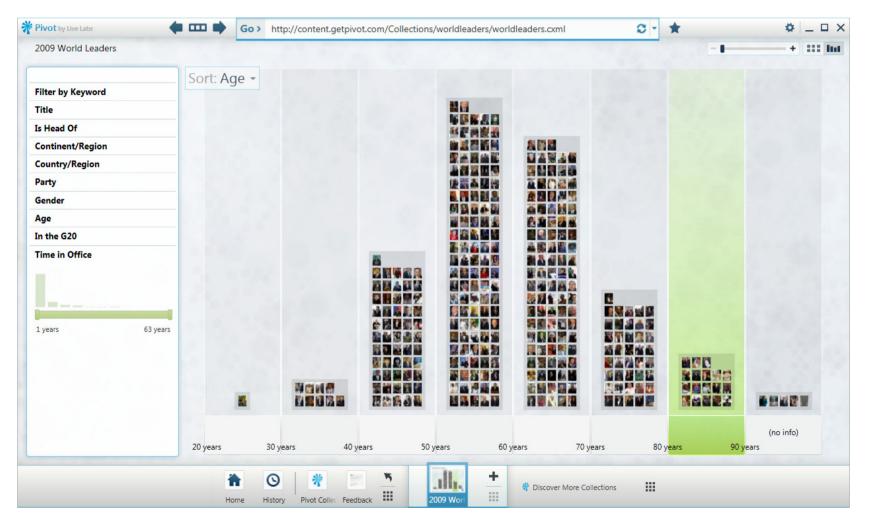
- SharePoint 2010
- IBM Websphere
- Windows Azure Table Storage
 & SQL Azure
- Zentity 2.0
- Services:
 - Facebook Insights
 - Netflix
 - Open Government Data Initiative
 - Open Science Data Initiative
 - DBPedia

OData Consumers

- Web Browsers
- Excel 2010
- LinQPad
- Client libraries for
 - Javascript
 - PHP
 - Java
 - Windows 7 Phone
 - iPhone (Objective C)
 - NET

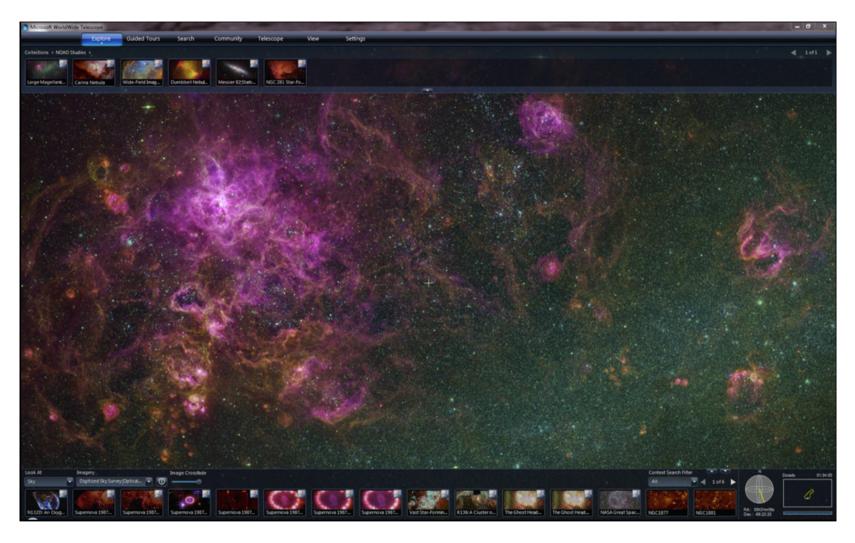






http://www.getpivot.com

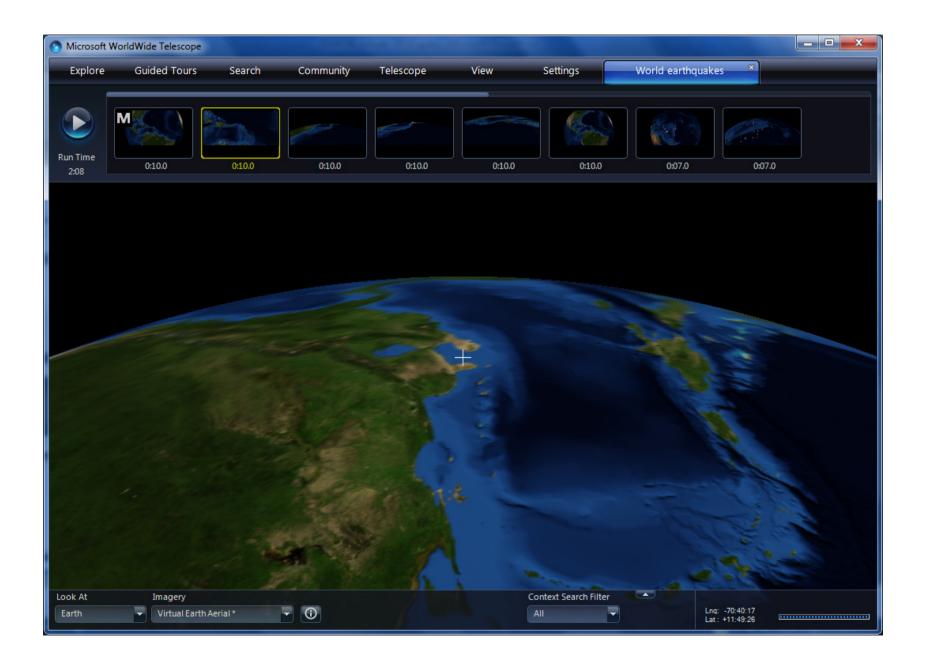
Worldwide Telescope

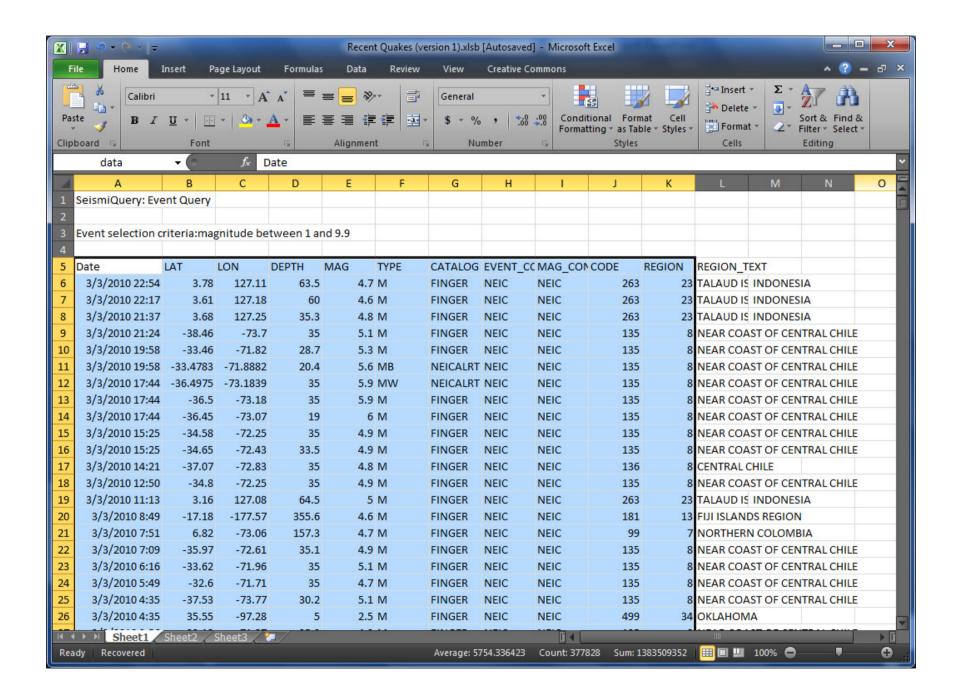


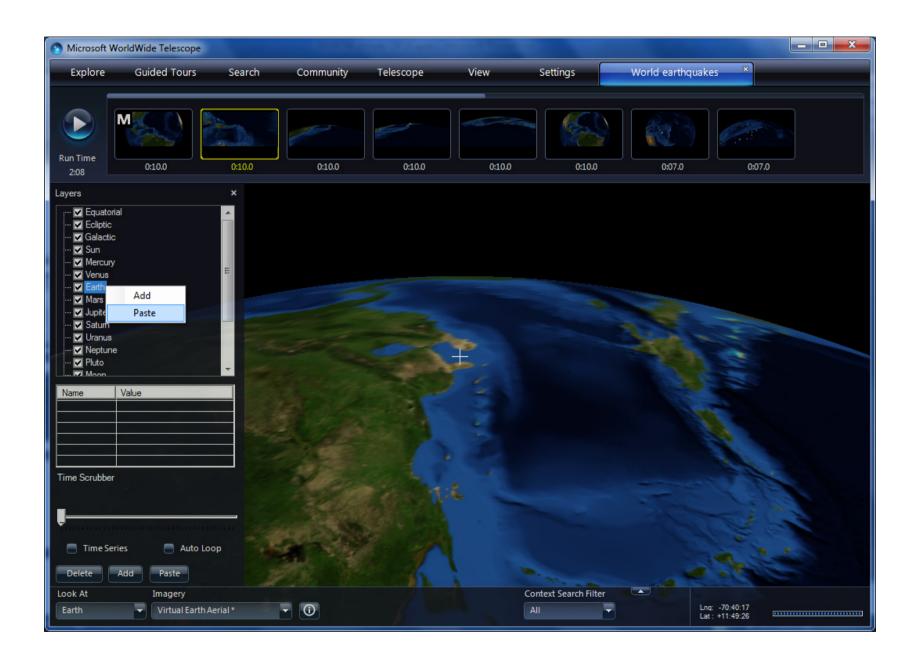
http://www.worldwidetelescope.org/



TIME magazine
"50 Best sites on the Internet 2009"









Tools for Repositories – Workshop *Join us on Friday! 2pm – 6:30pm*

Interested in new visualization interfaces, powerful semantic tools, and interoperability with Microsoft technologies to enhance your repository?

Program

14:00 – 14:15	Welcome & Overview Lee Dirks – Director, Education & Scholarly Communication, Microsoft Research
14:15 – 15:45	Zentity – A Research Output Repository Platform Featuring Pivot, OData, PowerShell, Excel/Power Pivot Oscar Naim – Sr. Research Program Manager, Microsoft Research
15:45 – 16:00	Programming with Zentity - SDK Oscar Naim
16:00 - 16:30	Break
16:30 – 17:00	Services for Repositories - RIC Alex Wade – Director, Scholarly Communication, Microsoft Research
17:00 – 18:15	Tools for Authors Alex Wade & Joe Townsend, Cambridge University
18:15 – 18:30	Wrap-up & Futures Discussion Lee Dirks

Questions?

Alex D. Wade

Director—Scholarly Communication Microsoft External Research

awade@microsoft.com

http://research.microsoft.com/people/awade

URL - http://www.microsoft.com/scholarlycomm/

Facebook: Scholarly Communication at Microsoft



Your potential. Our passion.™

