Shuffl: supporting curation of small-scale research data for web publication

1. This proposal seeks JISC funds to develop and trial an open source tool called Shuffl to assist small-scale research projects in the creation, organization and annotation of linked web data, facilitating preservation and eventual publication of those data.

2. The project will use as its theme a visual metaphor based on record cards, mimicking the use of physical cards in existing information handling tasks, to guide the creation of an electronic data management system with a user interface that is very approachable and easily adopted by researchers who are uninterested in the intricacies of data management, but whose work involves the creation of valuable data resources.

3. The intended outcome is a simple but useful application, supported by a lightweight back-end web service provisioned locally or through shared infrastructure. Data will be stored as RDF on a web server, accessible to other tools as linked web data.

4. The software will be open source, with the aim of seeding ongoing community development of a range of tools that build upon the underlying metaphor.

5. The project will run from mid-June 2009 to the end of November 2009, employing 60% of a single developer.

1. The data publication problem

6. These two responses were made at meetings we held with researchers to gather their views about requirements for data management systems:

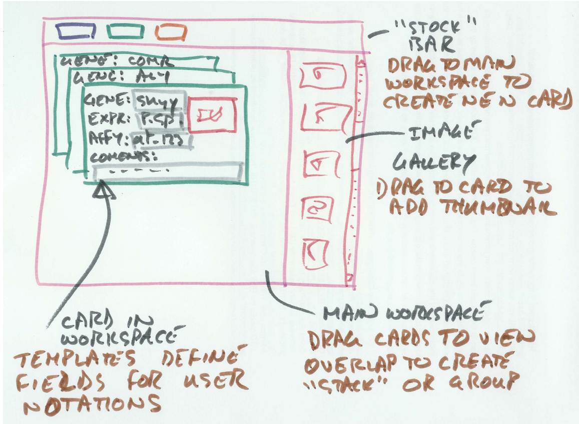
> *I'm in a small lab generating gene expression images by in situ hybridisation. I don't have the resources to develop an online database. I'd like something like Flickr which would give me a quick and easy way to share my images and maybe add some simple annotations.*
>
> *I know I should be organizing and archiving my data better, but it all seems too much effort.*

7. Our work with research data in institutional repositories [DefiningImageAccess] and with gene-function researchers [FlyTED], [FlyData], [FlyWeb] has exposed a fundamental impediment to the goals of data publication and sharing: unless research projects contain explicit funded provision for data management, acquired research datasets are maintained only to the extent that they serve to support the publication of research papers, and are then often allowed to rot. In contrast to high-throughput automated data acquisition systems that routinely have supporting data management facilities, small laboratory-based research groups often create data through expensive, labour intensive means, yet do not have the funding, systems, expertise or staff effort to support the organization and subsequent publication of their datasets, leading to loss of value disproportionate to their volume [ODIT].

2. Supporting data curation at source

8. We would like to provide small research groups with a straightforward and easy-to-use data management system that will meet them on their own ground, working alongside their existing methods of primary data capture, such as spreadsheets and imaging systems, facilitating creation of supporting metadata, enabling secure local storage with access for trusted collaborators, and reducing the effort of subsequent publication for re-use or submission to an institutional repository for long-term preservation. This will support workers in their day-to-day tasks, rather than impose unwelcome strictures, while opening a path for those activities to be migrated to a more automated environment, thus helping independent researchers and small research groups to capture their data in ways that are amenable to web publication, sharing and linking.

9. One application might be recording notes about genes to be studied: FlyTED [FlyTED] contains images and data on <1,000 from a total of ~14,000 *Drosophila* genes. Researchers' decisions to image certain genes may be for diverse reasons that cannot be recorded in a regular structure. Another possible application is illustrated below.


"STOCK" BAR
DRAG TO MAIN WORKSPACE TO CREATE NEW CARD
GENE: COMR
GENE: ALY
GENE: SHYY
EXPR: P.SP
AFFY: AT-123
COMENTS:
IMAGE GALLERY
DRAG TO CARD TO ADD THUMBNAIL
CARD IN WORKSPACE
TEMPLATES DEFINE FIELDS FOR USER NOTATIONS
MAIN WORKSPACE
DRAG CARDS TO VIEW
OVERLAP TO CREATE "STACK" OR GROUP

Putative image-annotation application

10. The first premise of this proposal is that effective data curation is achieved only when effective data management practices are part of the normal workflow of the researchers who really understand the data (a position supported by other commentators [PaulWalk]). Mechanisms to organize data in ways that support subsequent publication must also provide visible short-term benefits commensurate with the effort incurred. Immediate benefits provided by Shuffl will be a framework for organizing data according to researchers' developing perception, sharing of data and notes with colleagues, and (we hope) a user interface that is fun to use. And, to meet growing demands from funding bodies, simplifying data publication surely will have clear attractions for researchers.

11. Our second premise is that ideas and

methods used in general knowledge-working activities can be adapted to promote data sharing. Many day-to-day research data management activities are not so different from other knowledge-worker tasks, even if the goals and types of information handled are very different; e.g. consider the widespread use of spreadsheet software by researchers.

12. Our third premise is that users often do not start out with a full understanding of the patterns and structures in their data that will be important to capture, and that requirements will emerge as data collection and analysis proceed. Conventional data management systems tend to require definition of data structure in advance of collection of data, which must then conform to that structure. But real life often isn't like that, with details of important structures emerging only after some data have been collected. Our aim is to permit collection of data in a sharable form, and apply structure later, as needed.

3. Project description - creating web data through a record card metaphor

13. Drawing inspiration from successful pre-web systems [HyperCard], [NoteCards], and more specialized web applications [Mingle], we propose a system that adopts as a central metaphor the collection and evaluation of data in "bite-sized" chunks on record cards, where collected information recorded on each can be very loosely structured, or even completely free-form, linked to other electronic data, and stored together as a composite dataset.

14. In Shuffl, a record card will be used as a visual metaphor for a basic unit of data. Cards may carry arbitrary information, from free-form notes to highly structured data including links to arbitrary data such as spreadsheets and images. Many teams already communicate, brainstorm ideas and plan activities by writing on and arranging cards or other physical entities such as Post-it notes, since such methods specifically appeal to human perceptual abilities. The Shuffl system will enhance the obvious advantages of such physical systems by enabling the sharing of such card arrangements with physically distant collaborators, and by providing the security of electronic backup.

15. Other successful information systems have shown the value of permitting structure to developed as needed [MindMapping], [Idealist]. Shuffl will allow previously entered information to be organized and structured as useful patterns emerge, in two fundamental ways. Firstly, relationships between cards may be captured as typed links: part of the card metaphor's value is that it provides a finer level of granularity of information units than (say) document files, allowing these small units to be shuffled and reorganized to expose patterns in the data. Secondly, structure within the content of record cards may be developed using some combination of textual conventions, on-screen markup and annotation, and other mechanisms to be determined. This project will aim to develop these mechanisms in consultation with real research users, developing features that promise to deliver the most immediate benefit. An area to be explored is the interplay between research data collected in spreadsheets and a visual, direct-manipulation style of interface styled around record cards; e.g., each row of a spreadsheet might be treated as a card in a stack of similar cards, which can be spread out, clustered and rearranged in ways that aid discovery of patterns in the data. Another candidate for exploration is the combination of cards and images, where a displayed cards may contain thumbnail images and associated metadata; a visual interface could allow a user to group cards by appearance and then look for patterns in the metadata, or vice versa.

16. A weakness of this metaphor compared with physical record cards on a table is the limited scope for spreading cards out on a computer display for group interactions. In the immediate term, we hope the other advantages of a simple and direct interface will be of value. For the medium term, we expect to explore "fly-over" or similar mechanisms that make it easier to navigate a spread of more cards than can be seen in full detail at any time (the MacOS *Exposé* window-management interface suggests some promising user interfaces that might be explored). In the longer term, this kind of user interface could be extremely well suited to a forthcoming generation of large-format multi-touch displays from companies such as Perceptive Pixel [Ppixel] and Microsoft [MsoftSurface].

Applications

17. Although motivated as a platform for research data organization and management, this system's simple, unconstrained structure makes it amenable to use in a wide range of applications within and beyond the research community, including: lightweight ‘agile’ project management, knowledge elicitation and ontology design, user interface design involving ‘card sorting’ activities, annotation of images in image libraries, a virtual ‘semantic lightbox’, management of contact information, project tracking and scheduling, data provenance tracking, annotation of objects in collections, annotation of web resources, and the creation and organization of virtual collections, to name a few.

18. By using the web as the platform, and allowing access to data via simple web interfaces, we aim to encourage more specialized interfaces and applications to be developed around a core network of lightweight services.

Technical approach and standards

19. The user-facing application will be implemented using patterns employed in the

FlyWeb project for our OpenFlyData services [OpenFlyData], using browser-based XHTML, CSS and Javascript, with jQuery and RDFquery javascript libraries to look after the details of display interactions.

20. This front-end will be backed by a simple web service (e.g. a web server exposing a suitably access-constrained HTTP GET/PUT/POST interface, or a system like Caboto [Caboto]) to persist and share the note card data and associated data sets. Data from cards will be stored and accessed as RDF, available for publication as linked web data, and possibly also indexed using our implementation *[SPARQLite]* of the SPARQL query protocol [SPARQL] to support querying and searching. (This system could also work well in conjunction with a web-based file system, to simplify sharing of all kinds of data, but that is not a specific target for this limited project.)

21. Data lens ideas [Fresnel] will be reviewed and used as appropriate, as these represent an existing body of work for providing views onto RDF data. A possible area for further exploration will be the use of RDFa in displayable XHTML: one attractive option for crystallizing structure from free-form text is to add RDF annotations, which can be represented (and served) as standard web pages.

Other Web data authoring tools

22. Similar lightweight web-based approaches to data authoring include Semantic wikis [SemWiki] and web annotation systems [Annotea]. Shuffl distinguishes itself from these systems though a user-first focus on a visual interface for manipulating and arranging data, with back-end data integration treated as secondary (albeit vital) concerns in its design. There is also a fine-grained RDF authoring system in development at MIT [Tabulator-redux], but this requires a relatively sophisticated supporting infrastructure.

4. Fit to the JISC call

23. This proposal directly addresses two of the priority areas highlighted in the JISCRI call: Semantic web / linked data, and user interface design. The system is, at heart, a user-approachable way to create linkable (RDF) data, along with links to existing resources such as images and other data. The technical underpinning will include a very lightweight and generic shareable infrastructure service, and as such will help to demonstrate what can be achieved in this fashion.

24. The proposed system indirectly supports open data mashups and data search with mechanisms to deliver research data in a form that can be accessed by such services. It takes a novel approach to personalization in promoting mechanisms that allow users to evolve their own information structures appropriate to their research needs.

Alignment with existing JISC projects

25. This work is, in various ways, a continuation of the JISC-funded Defining Image Access and FlyWeb projects. Defining Image Access highlighted the paucity of research data suitably structured for preservation and publication, and sensitized us to users' resistance to engage in the complex processes needed to achieve these aims; it was also an early example of a JISC project that committed from the outset to progress records being made open and accessible through a public wiki. The technical approach for Shuffle (a modular, browser-based user application coupled with a simple, application-agnostic server back-end) builds on patterns and software developed by the FlyWeb project; furthermore Shuffl might become an active component in FlyWeb's OpenFlyData framework.

26. A small, flexible system with minimal systems or ontological commitment should be able to work with a wide variety of web-based systems. As such, we believe our approach is entirely in accord with the zeitgeist that appears to inform many current JISC developments in the research and teaching information environments. Our purpose is to provide useful functionality without getting in the way, and to support web-based integration of resources. Beyond that, the system aims to be largely agnostic about how it is used and integrated with other systems.

5. Quality and robustness of workplan

Project management

27. Project management will be based on Agile methods, organized into 2-3 week ‘sprints’, using a public wiki to record long-term milestones and goals, sprint plans and progress against these plans. At some point in the project, we hope to be able to use Shuffl itself, rather than physical cards, to record and manage aspects of the project planning (e.g. user stories and task breakdowns).

Quality plan

28. A cornerstone of agile development is automated testing: new functionality is implemented to satisfy test cases, which become part of the software's ongoing quality maintenance. Lessons learned from FlyWeb include patterns that maximize the unit test coverage of Javascript code, and use of the Selenium integrated development environment for automated user interface testing.

29. Documentation and other materials will be publicly available for review by users and peers.

Risk analysis

30. *Technical risk.* We do not aim to create a final system, but one that is useful, stable and can form the basis for ongoing community development. The technical risks are mitigated mainly by aiming to create a fully usable system with limited functionality within two months, and then adding functionality incrementally, always maintaining a working system at the end of each sprint. Then,

Table 1: Project plan outline

The project will involve three parallel strands: user interface development, back-end supporting service development, and user community engagement.

Month	Activity	Deliverables
1	Initial planning. Project setup, and further discussion with JISC OSS-Watch about open source tooling, licensing and governance issues. Prototype user interface in Javascript.	Project plan. Publicly hosted source and project tracking facility. Interface prototype that can be shown to users to elicit feedback and further requirements.
2	Develop supporting back-end service, and canvass potential users for feedback on initial interface - identification of a minimum system that target users would actually use for some purpose, however trivial.	Initial front-to-back system. Basic user documentation. A "hit list" of desired features.
3	Enhancements to system; capturing structure in data; basic user authentication and access control. Use of system for project tasks. Seminar/demonstrations to solicit users and feedback.	Updated system. Examples of system in use. Updated feature list.
4	Further enhancements. External linking and data integration. Continued user engagement.	Updated system, more examples of system in use. Updated feature list. Initial community web page describing the system and applications.
5	Create promotional materials (webcast, dissemination, demonstrations, etc.). Establish a long-term home for the demo service. Ensure that users with whom we have engaged are comfortable with what is available as this project winds down.	Public repository contains fully tested software, documentation, promotional materials, deployment instructions, examples, etc., in a state suitable to support ongoing open community development.

even if progress is less than anticipated, the risk of ending the project with nothing at all is very small.

31. *Staff recruitment and retention.* The staff member is already in place, so there is no recruitment risk. Loss of this key staff member would severely impact the project, but is unlikely due to its short duration. If it occurred during such a short project, effective mitigation would be difficult and uneconomic. Open sourcing of project materials from the outset mitigates the risk of losing all outputs in such an event.

32. *Failure to engage users.* Appropriate direction of our developments is predicated on engagement with potential research users. If we are unable to attract users' interest, outputs may be correspondingly impoverished. However, we are embedded in a research department, so potentially have access to a wide pool of researchers, some of whom have already volunteered to be our test users. The named developer is also working on a parallel data web deployment for integrating classical arts data, and thus will have opportunities to engage humanities researchers, adding value to that development. Non-research uses for the Shuffl system will offer alternative ways to explore how this development can offer visible user benefits.

6. Engagement and dissemination

33. Community engagement is, in principle, central to this development, though necessarily limited in scope by the nature and duration of the project. We will attempt to exploit existing contacts among biological and humanities researchers to get early feedback and user interest.

34. By starting with a small, simple system, we aim for early engagement with users for small tasks, and to build out from there. As soon as a minimal system is available (and stable), we will solicit feedback from colleagues and researchers in Zoology, Classics and e-Research. Contacts who have already agreed to work with us include:

35. *Jun Zhao*, EPSRC Life Science Interface Research Fellow and former researcher on the Defining Image Access and FlyWeb projects, who is already collaborating with biological researchers within the department and elsewhere to develop the use of linked web data in support of life science research.

36. *Professor Donna Kurtz*, Beazley Archivist in the Classics Department at Oxford, with whom we are working on re-purposing software developed for OpenFlyData to support classical art research.

37. *Dr Helen White-Cooper*, Cardiff University, with whom we continue to collaborate on *Drosophila* data projects.

38. *Dr Chris Holland*, Department of Zoology, Oxford, who is researching silk and other biological polymers.

39. We will use seminars and demonstrations to disseminate the work and gain further feedback.

Table 2: Shuffl project budget

	Year 1	Year 2	Total
Oxford University: Directly incurred staff			
Research officer: Graham Klyne, IBRG			
Oxford University: Non-staff			
Travel and expenses			
Equipment			
Total directly incurred non-staff (B)			
Total directly total (C = A+B)			
Oxford University: Directly allocated			
Staff: Principal investigator: Dr David Shotton, IBRG, 5%			
Estates costs			
Infrastructure technician (centrally levied for laboratory-based departments)			
Directly allocated total (D)			
Indirect costs (E)			
Total project costs (F = C+D+E)			
AMOUNT REQUESTED FROM THE JISC (80% FEC)			
OXFORD UNIVERSITY CONTRIBUTION (20% FEC)			

7. Budget and sustainability

40. *See table 2.* Non-staff budget provides for a portable development/demonstration system capable of running server, client and development software, travel for relevant conferences and visiting potential collaborators, and hosting for a public demonstration system for the project's duration.

41. The simplicity and flexibility of this system, coupled with immediate value for users, makes it an ideal candidate for an open source project that may attract a community of both users and developers. Source code will be maintained in one of the major Open Source repositories (e.g. Google Code).

8. Project team

42. David Shotton, PI, is head of the Image Bioinformatics Research Group, and has successfully led our two previous JISC-funded projects, Defining Image Access and FlyWeb.

43. Graham Klyne is the designated developer for this project. He has worked on IETF and Web standards development, and is co-editor of one of the current RDF specifications [RDFConcepts]. He was project manager for the Defining Image Access project and consultant and part-time developer for FlyWeb. Currently, he is working with the Oxford e-Research Centre and the Oxford Classics Department to apply the OpenFlyData software and experiences to multi-site integration of classical art data.

9. References

[Annotea] – Annotea project: W3C web annotation system, http://www.w3.org/2001/Annotea/

[Caboto] – Annotation storage for 3 JISC projects - http://code.google.com/p/caboto/

[DefiningImageAccess] - http://www.jisc.ac.uk/whatwedo/programmes/reppres/definingimageaccess.aspx

[FlyData] - Laboratory data management and decision support - http://imageweb.zoo.ox.ac.uk/wiki/index.php/FlyData_project

[FlyTED] - http://www.fly-ted.org/

[FlyWeb] - Linking Laboratory Image Data with Public Databases and Publication Repositories, http://www.jisc.ac.uk/whatwedo/programmes/resourcediscovery/flyweb.aspx

[Fresnel] - RDF data lenses, http://www.w3.org/2005/04/fresnel-info/, http://www.w3.org/2005/04/fresnel-info/fsl/

[HyperCard] - http://en.wikipedia.org/wiki/HyperCard/

[Idealist] - Idealist text indexing database http://www.chr.org.uk/idealist.htm

[LDOW] - http://en.wikipedia.org/wiki/Linked_Data

[MindMapping] - Mind mapping systems http://en.wikipedia.org/wiki/Mind_map

[Mingle] - http://studios.thoughtworks.com/mingle-agile-project-management

[MsoftSurface] - Microsoft Surface, http://www.microsoft.com/SURFACE/Default.aspx

[NoteCards] - http://en.wikipedia.org/wiki/NoteCards

[ODIT] – Oxford University, Office of the Director of IT: Scoping digital repository services for research data management - findings http://www.ict.ox.ac.uk/odit/projects/digitalrepository/findings.xml

[OpenFlyData] - http://openflydata.org/

[PaulWalk] – Repository architecture principles at http://blog.paulwalk.net/2008/07/07/repository-architecture-83/

[Ppixel] - Perceptive Pixel multitouch display screens, http://www.perceptivepixel.com/

[RDFConcepts] - http://www.w3.org/TR/rdf-concepts/

[SemWiki] - Semantic Media Wiki - http://semantic-mediawiki.org/wiki/Semantic_MediaWiki

[SPARQL] – http://www.w3.org/TR/rdf-sparql-query/, http://www.w3.org/TR/rdf-sparql-protocol/

[SPARQLite] - http://sparqlite.googlecode.com

[Tabulator-redux] - http://events.linkeddata.org/ldow2008/papers/11-berners-lee-hollenbach-tabulator-redux.pdf