The Future of DataUp

Hosted by National Center for Ecological Analysis & Synthesis, 735 State St, Santa Barbara CA

Relevant Links

Main website: dataup.cdlib.org
DataUp features: dataup.cdlib.org/dataup_features.html
Web app: dataup.org
Code repository: bitbucket.org/dataup/main
List of improvements: bitbucket.org/dataup/main/wiki/improvements_issues
Original requirements: Requirements.pdf
DataUp EML (google doc)

Meeting Goals

Funding for DataUp development ended in 2012. This meeting was intended to gather stakeholders to help to chart a course for where DataUp should go next. Meeting Goals:

1. Identify groups likely to benefit from DataUp
2. Evaluate and perhaps redefine the overall structure of DataUp (add-in vs web service only vs some other model)
3. Determine high priority development tasks
4. Identify potential DataUp contributors and community members
5. Establish goals for development given different funding scenarios

---

Agenda

Tues 2 Apr

9:00 - 9:45	Welcome & introductions
9:45 - 10:30	DataUp project & tool overview
10:30 - 11:00	Break
11:15 - 12:00	DataUp technical overview & development constraints
12:00 – 1:30	Lunch
1:30 - 2:45	Stakeholders discussion
2:45 - 3:00	Break
3:00 - 3:45	Features brainstorm
3:45 – 4:15	Features prioritization
4:15 - 4:30	Final thoughts; goals for tomorrow
4:30 - 6:30	Break
6:30	Group dinner

Wed 3 Apr

9:00 - 9:30	Summary of yesterday & goals for today
9:30 - 10:30	Your input & stakeholders
10:30 - 10:45	Break
10:45 - 11:15	DataUp governance: now and future
11:15 - 12:00	Prioritize development
12:00 - 12:30	Future directions & involvement
12:45	Group lunch
2:00	Adjourn

Participants

Amber Budden	DataONE	aebudden@dataone.unm.edu
Ben Leinfelder	NCEAS, DataONE	leinfelder@nceas.ucsb.edu
Bryan Heidorn	University of Arizona	pbryan.heidorn@gmail.com
Carly Strasser	California Digital Library	carly.strasser@ucop.edu
Chris Lortie	York University	lortie@yorku.ca
Dave Vieglais	DataONE	dave.vieglais@gmail.com
Eric Schultz	OuterCurve Foundation	eschultz@outercurve.org
Jeff Gerbracht	Cornell University	jag73@cornell.edu
Jim Regetz	NCEAS	regetz@nceas.ucsb.edu
Lisa Federer	UCLA	lmfederer@library.ucla.edu
Mark Schildhauer	NCEAS	schild@nceas.ucsb.edu
Robert Waltz	University of Tennessee	rwaltz@utk.edu
Sarah Clark	NCEAS	saclarky@gmail.com

Meeting Notes: Day One

Introductions & Background

Attendees represented a range of backgrounds and current disciplines; we were interested in getting feedback on DataUp that would help inform future development. Some attendees were well-versed in the DataUp tool, while others were fairly new and/or had not tested out the tool's functionality. In order to ensure attendees were starting from the same level of understanding, we spent the first half of day one providing a project overview, explaining the origins of the DataUp project, and reviewing the tool's functionality.

We first discussed the project's origins, including the collaboration among the Gordon and Betty Moore Foundation, Microsoft Research, and the California Digital Library. We then reviewed the background research that was performed during requirements gathering, and in turn informed development. We spent quite a bit of time discussing the decision made during the project to develop both an add-in and a web-based application. There was discussion about what audiences are served by the two versions of DataUp.

We then reviewed the current DataUp tool (both add-in and web application). We examined each of the four major features of the tool in detail: (1) Perform a best practices check, (2) generate metadata (EML), (3) generate an identifier and data citation, and (4) post data to a repository.

We discussed the process of creating a core set of EML elements to meet the needs of the DataUp tool. We also discussed the different types of identifiers available for datasets, why we chose ARK identifiers for DataUp and ONEShare, and explained the relationships between EZID, DataUp, ONEShare. In a similar vein, we spent quite a bit of time explaining the nuances of the ONEShare repository and DataUp's relationship to DataONE.

After lunch, we explored the technical details of the DataUp web service, the web app, and the add-in,

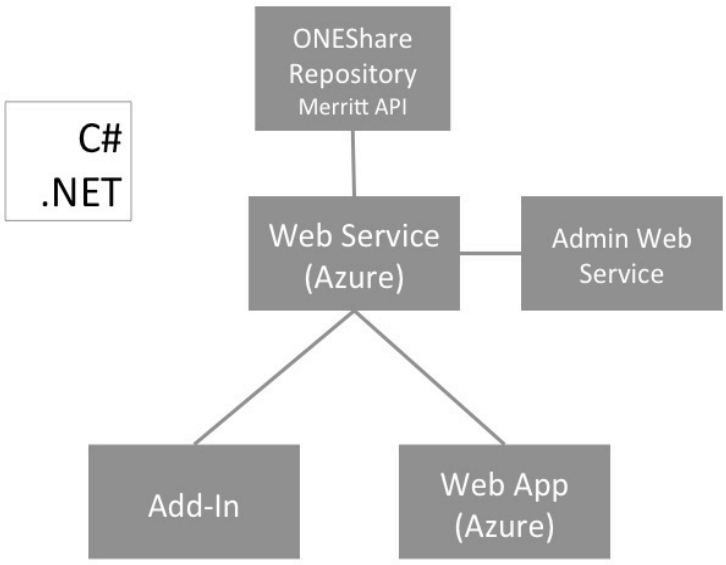

Figure 1: Current technical structure of DataUp & ONEShare

as well as the ONEShare repository and DataONE. We had a lengthy discussion about the status of the DataUp code, especially whether C#/.NET was a logical choice for future development of DataUp.

We examined the list of suggested improvements and issues available on the DataUp Bitbucket website[1], agreeing that there was much that could be done to make DataUp a more useful and usable tool for researchers. However we then discussed the constraints currently at work for making these changes: no one at CDL or DataONE has knowledge of the current code base, no one has C#/.NET skills, no one has experience with Visual Basic (the add-in code language), and we have no money to hire individuals that could help us.

After introductions were made and the project background was reviewed, we launched into a discussion of DataUp and where it should go next.

Discussion

Ben's opinion was that dealing with downloads, versions, and different platforms with their own idiosyncrasies. He's an advocate for developing the web app and dropping the add-in. He would rather work on a web app. We should try and enable it for google docs spreadsheets.

Robert suggested the plug-in be lightweight and minimal, and connect to the web service via REST API. DataUp is more extensible that way.

Eric said no one is working on the code right now, and we should therefore consider switching code languages now. It's reasonable to pick one that the community will adopt.

Amber pointed out that putting data online is still really scary to people. They aren't ready to share. Of course, they are using Dropbox and we could couch it in those terms. If we call it “backing up” your data, then it's a feature – we aren't “stealing” the data.

Bryan suggested that the web app should focus on the best practices check. This is the unique bit of the software.

Carly said that the strength of the tool was it removes some of the decision making: it tells you what repository, what metadata standard, what identifier, and what license you will use.

Mark and Jim wondered if we could leverage the KNB's ingest mechanism since it has a lot of the features we discussed. What DataUp does that Morpho/KNB ingest doesn't: best practices. What about a plugin to the KNB ingest tool? It could guide the user through editing and annotating spreadsheets?

Jim said we should focus on creating a specific type of tabular data: data tables that conform to the column/row format of a traditional database data table. We should constrain what they can put into the tables.

Ben said that NCEAS would like a better, feature rich way to submit data and metadata. It would be great to get Morpho/KNB to look more like DataUp. What if DataUp were a chain of components you could put together? E.g., quality checker + metadata entry + data deposit.

Lisa wondered how applicable our suggested changes and features would be to fields outside of Ecology. DataUp is generalizable right now. Will it stay this way if we make these changes?

Jim said that the metadata is generalizable. The EML could be mapped to other standards fairly easily. We could keep the tool in EML and then provide workflows for getting the metadata into other formats.

1 https://bitbucket.org/dataup/main/wiki/improvements_issues

Dave brought up the idea of integrating DataUp with shared folders, e.g., Dropbox, Google Drive, and Amazon Cloud. Most people are using these services already – might we leverage this to our advantage? It sidesteps one of the actions that the researcher must take and saves you from creating at least one user interface. (i.e., the upload interface).

Chris - For the best practices check, what about a "tracked changes" model similar to Flat File Checker? You could accept or reject changes?

Chris thinks our story could be: Dropbox + Share button + Flat File / Checker.

Other tools we could leverage:

Spreadsheet software to consider:

Google Docs, Excel, LibreOffice/Open Office formats

Software for checking the data:

LTER Quality Checker

DataUp

Geospatial, temporal, and taxonomy quality checkers

Flat File checker

Google Refine; Data Wrangler – more focused on facilitating the cleaning process.

Software for metadata creation:

Morpho/KNB Ingest

DataONE list: http://www.dataone.org/search/node/metadata%20type%3Asoftware_tools

Summary of Day One from Carly

- What makes DataUp unique (and therefore features to focus on)
  - Links to Excel and more generally spreadsheets
  - Best Practices Check
  - Relatively discipline-agnostic
  - More useable than many existing tools (good UI)
  - Removes choices and guess work of data archiving.
- DataUp Target Audience
  - Primary audience is individual researchers with little or no knowledge of data management
  - Secondary audience is the repositories and organizations that support data stewardship for these researchers. If we can get their buy-in, then the researchers will follow.
- We should focus development efforts on the web application, not the add-in

- Interaction with DataUp via three potential portals
    - the web interface
    - a lightweight version of the add-in (ideally for different spreadsheet types)
    - a Dropbox-style folder interface
- We will keep the progressive/workflow structure of DataUp as it currently is, with modifications. The Workflow:
    - Checker for cosmetics + data structure (track changes style)
    - Metadata creation: fields pre-populated; header row ingested
    - Checker for metadata/data compatibility
    - Identifier creation and repository ingest

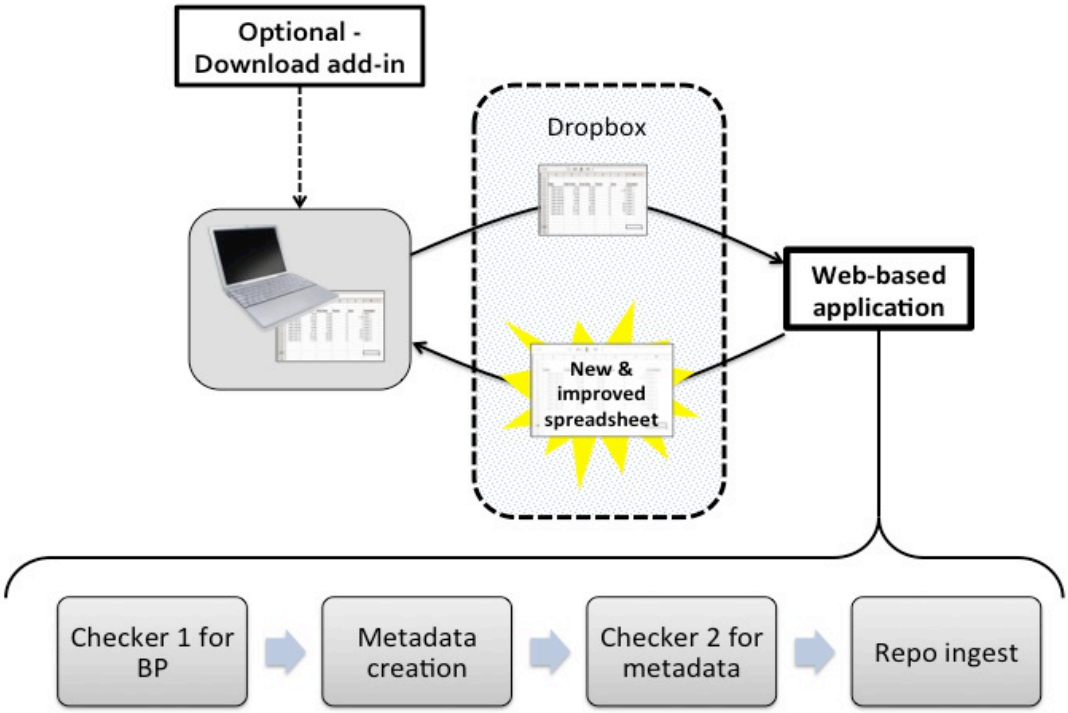

Figure 2: Suggested new structure of DataUp

Meeting Notes: Day Two

We began the day with a review of what we discussed the day before. We then continued Day One's discussions.

Add-in: What about Mac & PC Versions?

There were questions about whether an add-in for Mac was feasible; Eric told the group that it is a completely different language, and the current add-in would have to be written from scratch to accommodate the change. Essentially these two add-ins would be two separate projects, both separate from the web application. It was agreed that given the limited resources and time, we should focus on the web app.

The tension of DataUp

We want to help users format tables properly, ideally with adequate metadata. BUT some preservation is better than nothing. Dave: our target audience is people that don't know about data management. They are the simplest class of data generators.

Ideal workflow according to Dave

- Upload document
- Track changes to improve data structure and organization
- Approve or reject changes
- (metadata generation happens simultaneously with the two above)
- Download the file back to your machine

Note: we will keep the web app and the web service separate.

Possible Features:

- Ability to designate raw data versus more complex “working” data
- Ability to “suck in” the header row and allow the use to describe the attributes
- The data should be re-checked after metadata generation so that inconsistencies and problems can be identified. The best practices check and the metadata creator need to interact with one another.
- Pre-population of as many metadata fields as possible
- quality checking of geospatial, temporal, and taxonomic data (there are existing tools we could leverage)
- Use of controlled vocabularies, especially if we want to target a wide range of disciplines
- Ideally we would be able to support multiple metadata standards, not just EML (i.e., plug and play metadata)

- Track changes for best practices check
- It would be great to be able to highlight or otherwise designate your raw data. This way the messy, tabular Excel data could be archived, but the raw data would be identified by the author.
- We should try and keep the metadata tab. This enables metadata sharing.

What should be the goal of DataUp?

Amber: The end goal should be to share data and get it into a repository. It's important to remember that DataUp has an education component that helps people who aren't comfortable with data and metadata.

Bryan: DataUp has to improve preservation and access to a level that would satisfy funding agencies.

Ben: DataUp should be strictly online, focusing on checking data structure, describing the dataset, then sharing it. No mini-add-in or other fluff. This is a drain on resources with little reward.

Lisa: DataUp should remove the disincentives to sharing. DataUp removes barriers (like no knowledge of metadata, sharing, etc) so the researchers have an easier time taking care of their data.

Mark: I would like DataUp to be more proactive than reactive. Also three main features we should incorporate: header line ingest, track changes approach, and highlighting raw data.

Sarah: DataUp should hold the user's hand through checking their data structure.

Robert: The goal is to meet researchers where they work to encourage data preservation and metadata generation. By removing the add-in, we are no longer doing this. We should re-think dropping the add-in.

Jeff: Having the ability to check a spreadsheet against a pre-defined metadata set would be great. We should make sure that the errors come back in a way that the user can easily fix problems.

Eric: DataUp should be open source, fulfill a purpose, but is easy to modify/add on/ change etc. By encouraging data sharing, DataUp can be for data what what open source was for software.

Jim: Service should be guiding people not only to document and archive, but to use Excel better. Excel is NOT a data format unless it's used well. We should make sure we focus on tables versus worksheets.

Chris: the best practices feedback is most useful. DataUp should force a minor transformational shift in thinking for researchers. The goal should be to promote initial data sharing and basic metadata practices.