

Bluenote: Summary of the PlanetMath port effort after Summer 2011

Joe Corneli
September 22, 2011

Executive Summary: I describe stalled progress on basic aspects of the PlanetMath port, and then make a survey of the remaining issues that would take us to Beta. My ideal would be that we would go through these items as a team (i.e. I work on this full time and get help from other group members as needed) in order to complete the Beta by the end of October.

Porting PlanetMath into Planetary

The port of the old Noosphere-based PlanetMath.org to the new Planetary-based version has been progressing slowly. Some things are done, others are stalled, and a few other essential bits have yet to be tried. I will describe the state of affairs here.

As a brief account of how I spent my time over the summer: since I was running into trouble working on Drupal stuff, I thought about how convenient it would be to be able to hire a professional to just knock some of this stuff out, so I drafted a GRANT APPLICATION for JISC with that aim. I didn't manage to get it turned in (as I ran into difficulties during the team-building stage), but there will be a related funding round in October that I do think we'll be able to submit to (since I do now seem to have the team sorted out). Apart from this, I was first author on a couple PAPERS describing PLANS related to PlanetMath and MY PH. D.

But on-the-ground progress with coding was very limited. The three main topics I had hoped to deal with over the summer are (1) Metadata (like MSC codes for articles); (2) Permissions (Access Control Lists from the old PlanetMath); and (3) Problems and Solutions (with activity tracking). Constantin and I were also jointly hoping to finish (4) the autolinking service, which I'm hoping will be supplemented by (5) various text and interaction analysis tools to help students.

None of this has materialised, but I plan to talk to James Gardner *tonight* about the NNexus autolinker, and Constantin says he has also been making progress on his client-side autolinking tool. The basic aspects of the port (1-3) seem stuck. I'll describe outstanding issues for (1-5) in the following paragraphs.

(1) Metadata. According to our strategy, the canonical version of this information is stored in L^AT_EX headers, not in the database. It probably should be added into the database when the file is “processed for rendering”, since authors might change the metadata when they are working. Simultaneously, there may be a form-based way to edit this metadata, which should then be pumped back into the source documents. By contrast, Noosphere does, of course, store this data in the database. Injecting that old database content into Drupal should be easy. Then there's a question of implementing browsing over this (as we did with Virtuoso in the Vanilla prototype).

- There is also the matter of re-doing content transformations on a new dump of versioned articles from PlanetMath, since our old dump is somewhat out of date. I will talk to Aaron shortly about getting a new dump, and will re-run the transformations (getting Collab objects ported properly this time).
- Relatedly, Alpha is still not set up with LaTeXML rendering using local headers, which needs to be added.
- As a small related enhancement, it would be good to add *tagging support* at this point.

(2) Permissions. I installed the CONTENT ACCESS module, which allows per-article access control. Because some of the permissions are group-based, I tried installing the ORGANIC GROUPS Drupal module, but it didn't seem to provide what we were looking for, so I gave up.

(3) Problems and solutions In addition to being directly useful, this would prototype a Corrections feature for Planetary, since adding a Solution to a Problem is very similar to adding a Correction to an Article. There is a module for dealing with “NODE RELATIONSHIPS” that folks around OU advised me to use, but this isn't available in Drupal 7. So far I just added a Solution type, but a solution doesn't correspond to Problem (which is a problem).

- Note that once this sort of facility exists, the Corrections data still has to be ported.

(4) Autolinking service One of the bits needed by the autolinking service is some, ideally, meaningful URLs to link to. I installed the PATHAUTO module, and pages now have hyphen-separated URL aliases like *alpha.planetmath.org/abelian-group*. However, old links would tend to point to URLs based on the canonical name, viz., *planetmath.org/AbelianGroup*. This will all need to be aligned and made useful some how.

- I got NNexus installed, but not configured, and will talk to James about that shortly.
- Constantin will have some update on client-side autolinking too, I believe.

(5) Text- and interaction-analysis tools These things are not part of the port, but, rather, I hope they will be part of my contribution in my Ph. D. That said, I have yet to get into implementation work. My hope is to code up the Concept Forest [2] algorithm soon and make some experiments with that, to see if it works at all. If it does, I'm hoping we can plug some of the MathWordNet stuff in in the backend, and use this as a way to find “similar articles”. I'm also interested in:

- vocabulary acquisition by learners (so, want something to decompose texts to bag-of-words, and keep track of these in a history)
- I want to keep track of things like “helping others”, and some hooks into the underlying interaction model should allow me to keep track of that sort of thing.
- Finally, I'd like to do some analysis on use of heuristic reasoning in problem solving (which will partly involve stripping out the technical terms from the texts, and then doing some analysis and annotation of what remains). I also would like to introduce graphical “discourse markers” that can be used to explicitly indicate heuristic reasoning steps.

The question comes up: Given what we have achieved so far, how complete would this be as a port? Leaving aside (5), which will have an adequate proof of concept with NNexus integration, the the foregoing items plus the following items would get us through the Beta stage (if we don't try to do anything fancy).

(6) VCS integration Apparently we have basic commandline SVN integration in Planetary now, though I don't think there is a web-visible way to interact with that yet, and the corresponding VERSION CONTROL Drupal module is "not in Drupal 7". There may be other existing solutions besides that module (e.g. Drupal can manage revisions natively, but you have to ENFORCE this behaviour), but, in any case, integrating version histories remains to be done.

(7) User data Porting over a bit of user data into user profiles should be fairly easy, using the ADVANCED PROFILE and/or USER RELATIONSHIPS modules. Relatedly, in-site personal messaging data needs to be ported over (there is a PRIVATE MESSAGE module).

(8) ReCAPTCHA We are getting a bunch of spammy sign-ups so we should probably install RECAPTCHA or some other sort of obstacle to spurious sign-ups.

(9) Workflow We need some sort of "workflow" system, for which the RULES module seems the most sensible to use. (This would take care of things like adding points to a user's score when they add an entry, or saying that an article should be moved to the orphanage if it has outstanding corrections for 2 weeks, etc.)

Summary/Conclusion And I think that would do it. So, all in all it seems there isn't *that* much left, though as we saw over the summer, any one of these items can be stalled out and become a blocker. A few, like ReCAPTCHA are trivially easy (but also not that significantly important).

My ideal for the moment would be for us to go through this list in order as a team, since I'm personally stuck near the top of the list (we can skip (5) because I theoretically have plenty of time to work on that, and integration with Drupal will follow the patterns set up in integration of the LaTeXML and NNexus services).

I'm happy to take the lead on this (and will be putting in full-time hours to this end), but as we discussed last week, I do need some help! We may be able to get through many of the items by the time Constantin comes to visit me mid-October, and maybe all by the end of the visit. I have accordingly set the PlanetMath beta milestone due date for 21 October, 2011.

This would mean that any future *funded* work (i.e. in a JISC grant) could focus on feature building, rather than feature replication. A subsequent note will say more about that.

Bibliography

- [1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBPEDIA: A NUCLEUS FOR A WEB OF OPEN DATA. *The Semantic Web*, pages 722–735, 2007.

KEYWORDS: **semantic web, open data**

- [2] James Z. Wang and William Taylor. CONCEPT FOREST: A NEW ONTOLOGY-ASSISTED TEXT DOCUMENT SIMILARITY MEASUREMENT METHOD. In *WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 395–401, Washington, DC, USA, 2007. IEEE Computer Society.

ANNOTATION: I’ve been taken with the idea of using WordNet to determine document similarity quickly, and without the need for intensive offline analysis. Next step here is just to implement the algorithm and see how it performs. If it seems to behave reasonably, then a further step will be to supplement WordNet with a mathematical thesaurus (“MathWordNet”), using PlanetMath’s existing thesaurus. It would also be interesting to bring different-flavoured links into play; compare the DBPedia setting [1].

KEYWORDS: **document similarity**