CitEcCyr Work Plan

as edited by Thomas Krichel, with contributions by José Manuel Barrueco Cruz, Thomas Krichel and Sergey I. Parinov

Basic assumptions

CitEcCyr uses a network of two nodes, CitEc and SocioNet. The aggregation of the contents of both is called the common research information space. The SocioNet node will provide the CitEcCyr user interface.

Notation

TBD “to be done” […] is a reference to the flow chart.

SiP “Sergey Ivanovich Parinov” JoB “José Manuel Barrueco Cruz”

ToK “Thomas Krichel” ViL “Victor Lyapunov”

RoP “Roman Puzirev”

This text uses labelled tasks. All labels start with t\_. The only semantics in the label is the use of t\_o for optional tasks.

Tasks

Sysadmin -- t\_sys

t\_sys\_b A basic system set from bare o/s. This will use the same software as we have on the existing CitEc server, shuli. TBD by ToK for 1 week.

t\_sys\_p Document and port CitEc database tables. CitEcCyr will use the same database structure. If realistic, change some table and column names. In any case, document the database. TBD by JoB, 1 month.

t\_sys\_e Extend the database structure by data that comes in from the user services. TBD by JoB, SiP for 3 weeks.

t\_sys\_u Upgrade and maintain the server. TBD by ToK, ongoing.

Metadata ingest -- t\_min [B1]

t\_min\_r Formulate input metadata in ReDIF. ReDIF will be the basic input data from that the project supports. Status: mainly done, but changes will require some adjustments.

t\_min\_c Work on the conversion of incoming data to ReDIF. Status: to be done if new data arrives as on ongoing task.

t\_o\_min\_a Build a metadata ingest for AMF metadata into the table. This is a optional task that will only kick in if we have a source of data that will send us XML data that we have converted to AMF. ToK will work on this if required.

Fulltext ingest -- t\_fut

t\_fut\_s Develop full-text input in SocioNet. This builds a system where users can upload full-text files for a paper that has metadata in SocioNet. This is part of SocioNet Personal Zone. Status: done.

t\_o\_fut\_r Develop full-text input in CitEc. This builds a system where users can upload full-text files for a paper that has metadata in RePEc. This is an optional task that RePEc would carry out with the help of MPRA. Status: possible, but not actively planned.

Reference parsing traditional -- t\_rpt [B4]

t\_rpt\_s Build a sample set of Russian references in various styles. We need a few hundred citations hopefully from various styles that we encounter in the wild. TBD by SiP, 2 week

t\_rpt\_l Build a basic learning dataset for Russian citations. It involves manually parsing citations to delimit fields. The software should be able to act recursively, meaning that as we add citations and we add the reference as correctly parsed to the learning data. TBD by ToK for 2 months.

t\_rpt\_p Build a ParsCit customized installation that can parse reference strings into traditional metadata elements that ParsCit provides for. This is the core task of this work. ParsCit does not work out of the box with Russian citation. It has to be trained and may have to be extended. TBD by ToK for 2 months.

t\_o\_rpt\_e Build an extended learning dataset for known Russian entities, and discovered entities via a feedback. This would improve citation string parsing by offer a better training set. TBD by ToK for 3 months. This is an optional task.

t\_rpt\_p Publish training dataset at cyr.citec.repec.org. This will be done by ToK as soon as the first set is ready, but it may be a good idea for SiP to help with adding a description in Russian. TBD by ToK and SiP for 1 week.

Reference parsing XML -- t\_rpx [B4]

t\_rpx\_p Build a tool that parses XML references as we find them in the wild. We have XML references data in ReDIF in a matter that is largely undocumented. t\_dok\_x TBD by JoB in 2 months, probably spread across the project.

Citation contexts traditional -- t\_cct [B4]

t\_cct\_c As far as ParsCit allows, investigate all text files of CitEc documents in the text store for citation contexts. Build a simple XML store for this data, including some documentation. Status: TBD by ToK, 2 months

t\_cct\_s As far as ParsCit allows, investigate all text files of SocioNet documents in the text store for citation contexts. Build a simple XML store for this data. Status: TBD by ToK, 1 month, after t\_cct\_c

Landing page parsing -- t\_lap

t\_lap\_b Basic service for extracting the full-text from a landing page. The landing page is presumed to be available in the metadata where a PDF full-text link should have been provided. Status: partly done. JoB will improve, document and modularize the existing software for 3 months.

t\_lap\_c Checking of downloaded full-text objects. We need to do this in conjunction with t\_pdc\_s. To be done by JoB for two months.

t\_o\_lap\_e Build software that uses headless web browsers to download full-text from landing pages. This is an optional task. To be done by JoB and ToK if we have time at the end.

PDF conversion, t\_pdc [B3.2.1]

t\_pdc\_s Write a wrapper script around utilities to convert PDF to plain text. Status: mostly done, needs one more month to modularize, document, and build configuration options, TBD by JoB for 1 month.

t\_pdc\_h Write a wrapper script to convert PDF to HTML5 using modern tools. Status: TBD by RoP, 1 month, following similar output conventions as set out in task t\_pdc\_s

HTML5-based reference parsing -- t\_hrp

t\_o\_hrp\_p Build a system that parses references not from text, but from a HTML5 conversion of PDF files. This is an optional task TBD by SiP, ViL and RoP. This may take a long time.

Web service for editing references -- t\_wer [B6]

t\_wer\_s Write a specification on how the web service for editing references will look like. This will include the type of input, the actions that a user can perform on this input including how it will be exported in XML as user-supplied. TBD by SiP, 2 months. ToK will assist by pointing to inconsistencies and asking probing questions.

t\_wer\_i Implement the web editing service at SocioNet on SocioNet data. TBD by ViL and RoP, 4 months

Web service for editing citations, t\_wec [B8]

t\_wec\_s Write a specification on how the web service for editing citations will look like, including how it will be exported in XML TBD by SiP, 2 month. ToK will assist by pointing to inconsistencies and asking probing questions.

t\_wec\_i Implement the web service for editing citations at SocioNet on SocioNet data. TBD by ViL and RoP for 6 months

Citation linking -- t\_cil

t\_cil\_r Link parsed references in the database with papers to implement citation linking in CitEc. This is done, but we need to prepare the citation linking parts of CitEc into a separate software module that can operate on the assumed database structure. TBD by JoB and for 2 months.

t\_cil\_s Link parsed references in the database with papers to implement citation linking in SocioNet. This uses the software of task t\_cil\_r. This is done, but we need to prepare the citation linking parts of CitEc into a separate software module that can operate on the assumed database structure. TBD by RoP and JoB for 1 months.

Documentation -- t\_dok

t\_dok\_m Build a central documentation site at http://dicit.openlib.org that contains a description on how to redistribute data internally and to third parties. Status: done by ToK. This is first piece of real technical infrastructure done for CitEcCyr.

*t\_dok\_t Build a system to gather full-text and store them in archival files for citation processing and extended services. Lafka project, in progress by JoB and ToK.*

t\_doc\_i Firm up on the description of references and citations in both AMF and ReDIF metadata. Implement in ReDIF-Perl. Thus, this is not only a documentation task, but also involves changes to some complicated legacy software. Status TBD by ToK for 2 months.

t\_dok\_c Implement DiCit for RePEc data, by merging citation data from CitEc into the RePEc AMF dataset. This includes RelaxNG checking. This is [A2]. Status: TBD by ToK for 1 month.

*t\_dok\_f Build a system to document the current state of processing of documents from the databases. This includes a DiCit compatible format. Fraga project, by ToK, JoB, one month, in progress*

t\_dok\_s Implement DiCit for SocioNet data. This involves some software to be written to convert SocioNet data to AMF. While ToK has a ReDIF to AMF converter, this is about 15 years old and it is RePEc specific. Thus, this is a rather bulky task. This is [A1] Status: TBD by ViL and ToK for 2 months

t\_dok\_u Build a user input metadata format to export the data gathered from users in t\_fut, t\_wec\_i and t\_wer\_i. TBD by SiP and ToK for one month.

t\_dok\_x Document the XML structures we can parse from ReDIF metadata, see task t\_rpx\_p, by providing formal RelaxNG specs for them. TBD by ToK for 2 weeks.

Visualisation -- t\_vis

t\_vis\_s Build visualizations of the findings at SocioNet. TBD by SiP, ViL and RoP. This will be an ongoing task that will start when the first results are available.

Timetable, basic order of tasks

t\_doc\_t t\_doc\_f t\_lap\_b t\_rpt\_s t\_doc\_i t\_rpt\_l

t\_sys\_b t\_sys\_p t\_pdc\_s t\_sys\_e t\_rpx\_p t\_doc\_x

t\_vis\_s t\_doc\_c t\_pdc\_h t\_lap\_c

t\_doc\_s t\_rpt\_p t\_wer\_s t\_wer\_i

t\_o\_rpt\_e t\_cct\_r t\_cct\_s t\_wec\_s

t\_doc\_u t\_wec\_i t\_o\_fut\_r