Cost WG 5 meeting, Tuesday 16 Feb, Zagreb

Minutes, Neven Jovanović

AB = Arno Bosse
CvH = Charles van den Heuvel
GF = Gertjan Filarski
HH = Howard Hotson
IB = Ivan Boserup
LTH = Laura Tolosi Halacheva
MD = Mustafa Dogan
NJ = Neven Jovanović
NJef = Neil Jefferies
TS = Thomas Stäcker
TW = Thomas Wallnig

Arno's presentation

Will review most important elements of the call, less those already clearly described
starting communities = not yet received ESFRI funding
5 mil euros, min 3 partners, end of March - stage 2 not clear until August
stage 2 mid-March next year
signing of legal documents late 2017
funding starts Jan / Feb 2018
international partners can be included (esp. if technology unavailable in Europe)
Name of Call: infraia-02-2017
not for building new infrastructure, for integrating existing components
three types of activities: networking, transnational access, joint research - all three are mandatory
integration assumes starting community does not have much integration - we have an advantage here, because of the COST Action
strongest impact = networking (not just technical, also people and processes)
preservation, archiving, curation of data are considered relevant
needs to take into account ESFRI funded research infrastructure (most obviously DARIAH, CLARIN also others)
joint research activities: everything after accessing the data (virtual access)
industry is welcome in JRA, even expected
example for JRA: higher performance, integration of installations and infrastructures into virtual facilities (annotation)
innovative software solutions
what do we do with data past the point of discovery? - everything from here is jra
General & specific criteria
excellence, impact - the only things which will be evaluated during Stage I proposal

evaluation: excellence first, selection based on impact (weighted 1.5, but not in stage 1); threshold: 4 of 5
research & innovation: general; least generic, worth paying attention to: “beyond the current status quo” (meaning, beyond what is standard/wide-spread today)
excellence is also general: clarity, soundness, visibility
impact: reuse of material beyond research community, but also: whether the criteria stated are met
TW: what will happen at stage 2 in terms of evaluation? Requires details of work packages, plan of action, division of resources, management plan etc.? AB: Yes.

Phase 1 Template categories
Primarily the concept (judging excellence): objectives, relation to the challenge, concept / methodology, ambition (beyond the status quo), summary - table of partners, expected impacts
It is not always clear how the points are different, but it gives us flexibility to choose where we say something
18-20 pages in all, ca. 3-4 per section, plus cover, table.

N.B. We are not expected in Stage I to say in detail who is going to do what & with what (financial) resources; instead, focus on what we're going to do, with whom, to what end
The list is not final, we can add (even remove) partners for Stage II proposal as long as we don’t change the overall plan/concept.
We need to work faster - not rely on preparing the documents and present at the WG meetings
A coherent sketch (Google Doc) should be presented to the Action, then rewritten
HH - about the six points template categories and its redundancy: concept and methodology are very important, objectives can be brief, table is small, impacts follow clearly from the ambition; there is some breathing room
Methodology includes more than the infrastructure - also networking
Relation to the work programme - to the challenge
TW: “bring together”, “integrate” -- important words from the call
GF: no prior ESFRI funding - COST is not ESFRI
AB: only one project per research area will be funded - there are generic and specific needs; both focus and multidisciplinarity are required
GF: It would be good not to widen the scope too much, keep focused on the Rep of Let
TS: Where do we bring DARIAH and CLARIN in? Is some commitment required?
AB: Generally, there is nothing required. But, we need to reach out to them asap. Technical infrastructures will need to be integrated in some form. EUDAT, text mining, cloud-based capabilities could be explored as well but CLARIN and DARIAH have to be there
VU: what about outcomes, do they need to be mentioned in phase 1?
AB: Not in detail. The time between stage 1 and stage 2 - position papers on parts of infrastructure and parts of the bid, describing it in more detail, discuss in community, then print elaboration of what we want to do.
HH: We will use COST to bring out a volume about our aspirations (contributions to go online, commented on by community, then as a traditional publication)

AB: impacts in the call don't make a direct reference to outputs (effective research produces outputs)
TW: networking activity which opens dialogues between collections/scholars/developers can be a form of output (publication digital or analog)
MD: What is significance of not having a partner from industry?
HH: Worst case, it could be an opportunity to eliminate the proposal. Industry partners have to be brought in, both big and small.

Howard's presentation

HH - parts of our agenda map to application sections described by Arno; 4, 5, 6 will be discussed tomorrow. 5) Content providers are an important issue. 6) Process: how will this be done
Need to be decisive, structured and disciplined this time - we want to know how will we go ahead, with some firm decisions.
We start with a clear focus and potential for opening (chronology, geography, themes, textual reach, social dimension, material aspect exchanged in networks) - European cultural and intellectual exchange across ages
Do we mention the potential, for the sake of those who do not have an intuitive grasp of it?
We need to understand the mentality of the funding body. We have to ask people in Brussels (h2020 can tell us). National contact points are not authoritative enough here. Potential for expansion is attractive (a transnational entity existing for hundreds of years).
NJeff: Technology is not a limiting factor - content providers, community
HH: EHRI/Holocaust is a well-known entity, RoR not so. But Republic of letters is also a good metaphor for what we are doing with this action - transnational, from the bottom up. A demand-driven project.
TW: is the delegation to Brussels feasible before 30 March? Are the relevant people there (or in other parts of Europe)? The group-building is a slow process
HH: this belongs to tool-building agenda (prosopography is part of package); sustainability is also important to have in mind. Where to find the people? If somebody in h2020 would read and comment a mature proposal, this would be advantageous. It is important to ask whether our interpretations are right.
LauraT: a good infrastructure can be expanded, material part with more difficulty.
HH: Question of scale - 3.5 mil w/o overheads, 3 mil net; 1 for bringing in data, 2 divided among key infrastructure providers - needs to be prioritized. The core tasks which allow a fully functional basis for development and expansion.
HH: more about the scholarly community than technology infrastructure (which is relatively modestly developed)
AB: the money really should not be used for building something from scratch (capital funding is not allowed at all); we need to integrate what is already there; tools have to enter at (minimum) the prototype stage, use funding to attain production quality.
MD: governance? processes, workflows, data validation (cmp. Europeana) - this question can come from the commision
NJef - linked data can offer a way for somebody more restrictive to use what we have produced
Listing the partners who can help us

- which parts are existing
- which partners can help us assess the components
- which infrastructure components need to be developed
- who can help us develop

Neil's presentation

object model - data surrogates for object; detailed metadata, not completely standardized (hopefully being standardized); open-shareable metadata to be deduplicated; annotations (scholarly, transport layer for enrichment, analytic tools)
Ivan: relations between providers and the central point? The record has an id number elsewhere as well, how to show this?
NJef: depends on the collection holder!
HH: the modes of collaboration have to be clearly explained
IB: the idea is to use the project as a stimulus to raise the level of categorization
NJef: version history is possible in EULO in a box
The high-level ideas will be presented here, details of implementation later
AB: Benefit for the local institution / provider?
TW: the work of the developers/librarians is not free - neither is that of the scholars
HH: there is an incentive for institutions to join early on, when the risk is high
Laura: what about conflicting annotations?
NJef: they are attributed, they can be themselves be annotated. The knowledge model includes inconsistencies
APIs: for discovery, outside; for jra's, annotation
TS: an institution offers data, EULO promises enriched data back; what kind of format will be provided, RDF? German institutions are not able to harvest that data.
NJef: this is what is EULO in a box for
MD: content providers must use the feedback
TS: make a point: it is easy to integrate for your institution
NJef: network will provide assistance to show how this is done; outreach and enabling work is mostly human
HH: multiple visits to major institutions
GF: local integration must be possible
TW: a test case with small amount of data is also a possible approach
CvH: these are services - for experimentation (also training)
HH: one hour of time for a presentation, six months later we present something in an afternoon
NJef: four year cycle is a good opening
CvH: CLARIN provided time and expertise, it was a good approach
NJef: get together institutions and developers, this will be ok
IB: advantage for institutions - access local record, enter greater database from there; offering local users wider advantages
NJef (after break): Collection holder view of EULO in a box; scholars view
Leveraging existing infrastructure - a list is compiled
AB: vast majority of people are not in VIAF or the like -- there is a responsibility
GF: DBpedia?

NJef: can be on the list to be assessed. We would run a "sameAs" service.
TW: building new ontologies for Early Modern? (For social status of people -- possible technically, but hard to get it accepted)
IB: We want to show library users and librarians view as well.
NJef: This can also feed search engines etc.
CvH: there are existing factsheets of standards and tools...?
NJef: preservation service -- could be offered to partners (for their annotations)?
HH: place authorities were discussed in WG1 - they could provide expertise (by the end of March or April)
HH: we will be holding some data in the hub -- deduplication, enhancement; it can be propagated to all the resources listed. A historical gazetteer of Europe will be incrementally produced by scholars all over Europe. Such authority files could be major outputs of the project.
CvH: this is tempting, but how to organize it? What to do in cases of conflict?
TS: enrich already existing agencies is a good point (when we identify letters, we are the source)
HH: sources for identifications -- potentially, our contribution is large, more people wrote letters than books
MD: persistent identifiers for letters link to DARIAH, it could be of interest for them and for us
HH: how to assess just one candidate? What assessment does NJef need? WG 1 can provide this for place and time authorities. Biographical resources: wg 2, place/time: wg1
NJef: this leaves open identity / authentication, identifiers, hosting, preservation; we can say that we are considering
AB: but we also have to say we're integrating (provide list, assess)
TS: criteria -- quality of data, reliability, accessibility of data
GF presentation
Implementation of an infrastructure has already started at Huygens, but what will it do? -- Illustration of workflow. It starts at the analysis stage (materials are digitized and transcribed)... but it cannot start there, we have to take into account transcription etc.
Annotation: Huygens - Alexandria store
Resolution - Timbuctoo store (knowledge graph -- important also for parsers)
Example: Resolutions of States General - very structured, very varied corpus
A very large, very slow project - cannot be accomplished in human history
Transkribus can recognize handwritten text -- it should be looked at!
This is a good chance for integration
TW makes a case for using transcription to train students
HH this is something that any scholar can understand and relate to
Analysis -- semantic parsing, language recognition, resolution of implicit dates, formulaic language
The number of annotations creates XML problem
Solution: switch to graph model
Problem: hard to scale up
Laura: compare technologies used ab ontotext
GF: RDF also does not scale well
MD: Level of granularity?
GF: token level (although it can go lower)

HH: two years before funding for four years -- what can be done in six IT years? We might be doing something else; we won't go this route necessarily -- we have partners involved in developing who could be solving a significant problem, integrating also emerging solutions. We would prefer state-of-the art solution, even though we have a fallback.
CvH: we can address the problem of technology advancing
TW: do we have to present a timetable?
AB: not at this stage, in the narrative only -- ambition includes being beyond the state of the art in the sense of beyond what is widely in use today
TW: glacial pace of scholars adapting to technologies - timeliness of the application: if not now, it will diverge
GF: Resolution stage: different sources about same person / information / event
Within CLARIAH it will be more generic -- currently limited to what Huygens scholars need
NJef Follow the graph and come back to library (to enrich it)
HH: tools and analysis -- what about analysis? Priorities for tools?
TW: Definition of “Joint research activities”?
AB: No single definition, but it can be thought of as ‘everything you do with data once you get an access to it’.
CvH: Interface for queries, visualizations
Laura: a combination of traditional and innovative ways
NJef: in Oxford (and at Huygens) there are different interfaces
HH: what tools do we want to integrate? What partners do we want to get to do it?
People are exchanging user histories: find me everything by person X... send it to Transkribus...
GF: can we feed Transkribus annotations an annotation schema?
TS: A better approach -- define interfaces, define use cases
AB: This has to be defined for EULO in a box, and it has to be done in Zagreb!
NJef has to show the structure
1 Fedora as a versioning object store
2 hydra - blacklight for basic discovery
3 interfaces to feed into the system
4 editorial workflow (from EMLO)
5 linked data platform - RDF models
Who can contribute? Java / Groovy expertise - Huygens can do that
GF: how does this stack relate to other choices? Has to be discussed. Huygens would take on a part of it.
Vertical organisation between 2-3 partners, test-driven fashion
key expertise: integrating with another infrastructure
Laura: technical issues she needs to understand, they have a triple store
NJef: another triple store can be plugged into fedora
MD: they would be interested in minting identifiers
AB: Stanford needs to be reached out to now; Gertjan can help us come up with certain criteria to structure the decision...
AB: for technical partners another revision is needed, with a bit more context; for content partners this is enough
NJef - possibility of collaboration with Stanford (people, probably for free)
GF: can offer three to four people, Gottingen as well - 12 engineers

AB: what expertise is needed?
GF: people pick up technology quickly, but to get them is a problem
NJef: Stanford can give annotation engine (hypothes.is), Microsoft would offer
MD: DARIAH could offer hosting
AB: it would be good to sketch up requirements
GF: there are risks in putting people in phase 1, they will expect something..
HH: is it better to say "with a small group we will accomplish..."
NJef: Stanford in advisory capacity
GF: three major technological partners would be ideal
TS: if three partners are minimum, would this open up place for a commercial partner?
HH: one small to medium enterprise industrial partner would be welcome
HH: A good powerpoint presentation for data partners March 4 is needed - recruitment
AB: A look at the calendars is needed for availability.
MD: and some time for internal processes is needed
GF: for the next round the financing and the whole process will be needed, for now it's ok.

Wed 17 Feb

Thomas' presentation

TW Historiographical agenda
networking and historiography
the call and our proposal
networking activities
COST has a technical agenda and a historiographical one (researchers, not content providers: editors of letters, commentators...)
WG 2, 3, 4 connect technical points with theory
the term "methodology" is used in a different way in historiography (vs. technical side/bid)
translate functionalities (visualisation, statistics) into existing scholarly language: history of knowledge, Begriffsgeschichte
Showcase a broad portfolio of historiographical resources, from traditional to network / social context to document-oriented archival approach: not only one need, but a broad spectrum we're addressing
HH: explains tasks of wg's and theoretical layers: align workgroups to well-established practices in the discipline, as a set of complementary approaches
HH: the application will be read by IT specialists, the descriptions have to be intelligible to non-specialists
TW: if somebody is sent to Warsaw to assess value for community, we should be prepared
CvH mentions the Vienna document which should be revised in light of what TW is saying now
TW: introductory sessions should be shaped like this, the rest can deal with case studies
IB: where is media research?
CvH: also history of collecting, material culture
TW: call about networking activities -
pooling of distributed resources is the core of what we are doing

CvH: not only monograph, but also enhanced publication (with website) - a networking activity
AB: commented online articles demonstrate the understanding of the community
CvH: involvement with a publishing house? enhanced publication - data, comments goes over to media
AB: joint management - for reviewing contributions; curation quality standard?
MD: automated validation could be possible
TS: is the project in position to enrich or convert data?
HH: data in whatever form would be welcome, because providers were not able to do it on their own; just knowing that the letter exists is a start. Non-standard data can be excluded if necessary
NJ joint management: how to apply for EULO in a box, updates, help?
TW common standards are connected with joint research activities (authority, ontologies)
EULO in a box roadshow: spreading good practices, outreach and training; also STSM - use the tools which were tested in cOst action - here ends the excursus discussion that in fact pertains to the Joint Research Infrastructure part
TW: for stage 1 proposal
types of activity

- STSM, training, workshops

types of scope / audiences

- scholars, collections
- put money into smaller collections
- test data merging

HH satisfied with STSM and training schools (young energetic people); training workshop is different from roadshow
HH issue of publication - not traditional article in traditional journal; highly innovative publication plans addressing the project core, keeping at bay traditional requests
TW does not think this is a big danger; scholars, librarians and developers all have to get funded
HH the call is not about the discipline, about infrastructure; allow scholars to graft research projects onto infrastructure (funding to help people write grant proposals); experience from CofK - help people formulate their project
AB DARIAH tried to set up virtual competency centers
MD: could the resources be shared? Would there be opposition?
HH does not think so; application reviewing is hard work
AB mentions US models of humanities centers (UV Scholars' Lab - serving graduate students and less faculty, training them to get the kind of competence they need)
AB will open the Google Doc, keep it until 25 Feb
HH is 500.000 enough? OK? (everybody agrees) - a website will also be required, can raise the sum if necessary
AB about name EULO - reassembling the republic of letters (people think it a good idea)

Laura's presentation

LTH presentation of ontotext - a company dealing with heritage institutions
news enriched with concepts from dbpedia (based on open data repository)

TS to LTH: how does ontotext provide provenance of information?
AB why is it 85% likelihood in machine inference?
NJef ontotext could use existing texts to make inferences for an amount of data cleanup; also enrich data -- this can be offered back to content provider
HH to LTH - reconcile by accessing a domain-specific cloud of data
GF: can we use the ontotext knowledge base for other purposes?
LTH - the linked data is not property of ontotext, only the uri
TW - who will own the EHRI data?
LTH - probably not ontotext; she will get an answer on that
AB wants to know where ontotext makes money - database, consultation?
NJef also wants to know about pricing
GF - about open source software (regarding EU rules): where do they stand, can we get an open statement? (to compare open vs. closed)
TW - addresses balance between industry involvement and open source
AB would like to see contributions of ontotext for EHRI bid, and contact partners in other cultural institutions (why they chose ontotext, what were the advantages)
IB presentation of CERL
European library organisation delegated to CERL all issues regarding cultural heritage; projects distributed between Uppsala, Gottingen, London. The number of members (paying) is being expanded; there are also US members
CERL established a multi-lingual multi-cultural thesaurus and gateway, freely available and used - proquest is customer (EEB)
the database contains well-ordered synonyms of personal names
CERL portal of manuscripts (and early books)
2.5 million manuscript records, harvested from available files
they want to expand access possibilities - letters as a special group (access to Vienna corpus of letters)
there is a possibility of establishing a corporation
AB - what is status of CERL (legally)
IB -private company under English law (can be an industry partner)
HH - sees points of contacts and modes of interacting: authority file (eminent members of the RofL); would CERL be interested in a reciprocal relationship? Most urgent - with whom do we need to speak before the end of March?
IB - the secretary of CERL could arrange something without formalities (IB would suggest that this collaboration be for free)
CvH - this could be a joint research activity
MD - Gottingen uses CERL in research projects
HH - CERL portal of manuscript resources should also have an authority file for letter-writers; does CERL construct it? -- a quid pro quo. (This is acceptable) -- a letter / e-mail to secretary

Closing Discussion

HH chairing

- process of writing the application

- feedback from (major) contributors to the project (Herzog August, Royal Library, Prague)
- mix of contributors (large, small, different parts of Europe, archives, publishers, other data aggregators: CERL)
- how will the invitation to participate be received?

TS (on data sharing in Wolfenbuttel) - wg 4 works on that (identifying stakeholders); German national data points (Kalliope), but not clear on position in Europe; Dublin (Elisabeth Anne) is a potential partner. What expectations do the partners have? We will not be able to pay much for data, but we have to convince them that a win-win situation is (without funding). The project may trigger opportunities to bid for money at local funding agencies - use cases of partners working in tandem, contributing and applying locally for converting the data
HH would like to make this offer to a carefully selected group (card catalog, no electronic) - digitize, upgrade, standardize, ensure discoverability on an open platform; we will later offer a matching funding support. For March we need agreements with significant major institutional repositories.
TS wants to see hubs in countries -- finding collections, talking to people (examples of that kind listed for application)
HH - EULO on line as a national / regional repository
TW is working with Vienna in that direction, defining where do librarians work in the workflow
HH this means discovering hidden costs, and defining funding needs
TS: state clearly in the application that this does not make the national work superfluous
AB: treat digitization as enrichment, metadata as the baseline
TS: Europeana offering a contract in English was a problem - for administration
HH: attractiveness of national institutions - IB and ML
IB: The Royal Library is interested in displaying its material in as many ways as possible: outreach (Fabricius letters, MS letter collections) - they would like the long-term benefit of being included in technological and intellectual upgrading
HH - can a draft of letter, discussed with IB, be a way forward?
IB will first discuss it with the director; a flexible approach is needed - from paying something and doing all the work, to...
IB - their catalogue is integrated
HH specific letter collections make a lot of sense as a starting point, but creating catalogue from scratch would be expensive - upgrading an existing catalogue would be preferred
TS - how to identify the right letter collections? A list of persons would be a great start.
ML on hosting EULO in Prague, acting as a hub for Czech material, additional data to be fed into the project? -- Questions regarding technical issues. Scanning of catalogue cards -- ML does not know whether the Czechs need this and how they are solving it. A national node is reasonable; what space etc. should be set apart for data (and how to calculate costs)? Cooperation with other libraries: nationally, is it necessary? Sustainability - data repositories should be running after the project, have we thought about this?
HH - a list of questions is good.
ML - Academy of sciences does not function yet as a national hub for data (they host data for physics etc)
HH - would they be interested to function as a node for collecting material?
ML - yes

HH - Suggests ML sends questions for the project to collect (similar to ones from the Austrian National Library). No formal letter of agreement is necessary at this stage.
ML needs a description of what is required from a national node
HH asks them to prompt for more information. The process will begin in the next couple of weeks.
NJef will discuss this with ML in Prague.
HH -- within next two weeks we will be in touch
HH - the kind of data contributors: publishers (OUP - would like to integrate the Oxford Dictionary of National Bibliography, Droz; potential: Olschki); OUP is a candidate. Archilet (affiliated with a university) - what chance there is of them being affiliated with a permanent institution? Darwin correspondence in Cambridge: multiple overlapping scientific correspondences 19th c. (a thematic data aggregator). This could be a move beyond the RofL. Link to Cambridge University Digital Library.
TS: concerned about extending the corpus beyond Early Modern.
HH: they would do groundwork for their resource, but not mix with our repository
CvH A selection would be acceptable
TW: Expansion should be signalled, but with minimal cost; a training workshop in Cambridge would be a good thing (including the Darwin people)
AB: google doc with template (descriptions of topics), an outline in the next few days
MD: not on the distribution list, should be connected
Contributors: Bibliotheque Mazarine, Elisabeth Anne, Kalliope (rather difficult); Huygens as the national node; Leiden library can be contacted (outsourced digitization, beyond paywall, also for the catalog)
WE need a major French partner, a library tour might be in order.
It was agreed that Oxford, The Hague and Göttingen take on responsibility for the technical part of the project.