WP 2 - Central Hub

Dieter Van Uytvanck
CLARIN ERIC
dieter@clarin.eu
2015-09-09

CLARIN-PLUS Kick-off, Utrecht

Introducing CLARIN-PLUS

- Exceptional, one-time opportunity
- Contrary to many other H2020-projects this is a chance to be shamelessly selfish and to focus on making CLARIN better and more sustainable
- Mostly not about creating new things, but rather:
  - to get them into a better – more future-proof – shape
  - to integrate them more closely

General points

- Standardize deployment of server-side software:
  - Docker
- Move away from ad-hoc approaches/hacks
  - but take care
    - not to over-engineer alternatives
    - keep things manageable
- More sustainable solutions

Task overview

Task ID	Description	Responsible party	PM
2.1	Robust and extended SPF		
2.1.1	Extending the SPF coverage	CLARIN > Dieter	3
2.1.2	Robust technical foundations for the SPF	CUNI	10
2.2	Metadata quality improvement		
2.2.1	Metadata benchmarking and curation	CLARIN > Vienna	4
2.2.2	Improved harvesting workflow	CLARIN > Wroclaw	4
2.2.3	Reinforced concept registry	CLARIN > Meertens	6
2.3	Generic workflows		
2.3.1	Language Resource Switchboard	EKUT	6
2.3.2	Integration of new web services into workflows	EKUT	4
2.3.3	EUDAT service compatibility	EKUT	4
2.4	Infrastructure gateways		
2.4.1	(Federated) Content Search Engine	CLARIN > Språkbanken + IDS	5 + 3
2.4.2	Virtual Language Observatory	CLARIN > Twan	8
2.4.3	Virtual Collection Registry	CLARIN > Willem	8

ID	Task Name	2015				2016											2017							
		9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8
WP 1	**Project Management**																								
T 1.1	Project coordination												D												D
T 1.2	Administration												D												D
T 1.3	Technical coordination												D												D
WP 2	**Central Hub**																								
T 2.1	Robust and extended SPF																								
T 2.1.1	Extending the SPF coverage							M													D				
T 2.1.2	Robust technical foundations for the SPF										D													D	
T 2.2	Metadata quality improvement																								
T 2.2.1	Metadata benchmarking and curation						D																		
T 2.2.2	Improved harvesting workflow												D												
T 2.2.3	Reinforced concept registry			M													D								
T 2.3	Generic workflows																								
T 2.3.1	Language Resource Switchboard			M											D										
T 2.3.2	Integration of new web services into workflows												D												
T 2.3.3	EUDAT service compatibility																					D			
T 2.4	Infrastructure gateways																								
T 2.4.1	(Federated) Content Search Engine		M						M													D			
T 2.4.2	Virtual Language Observatory			M						M													D		
T 2.4.3	Virtual Collection Registry				M						M													D	
WP 3	**Central Office**																								
T 3.1	Enhanced Intranet								D																
T 3.2	Intranet: content enrichment																				D				
T 3.3	Usability and website																								
T 3.3.1	Unified User Experience								D																
T 3.3.2	Language Resource Portal Integration												D												
T 3.3.3	Usability Testing and Guided Tour													D											
T 3.4	Site: improving navigation and content																			D					D
WP 4	**Infrastructure Partnerships**																								
T 4.1	Coordination with other infrastructures						M																		
T 4.2	Implementation of Joint Services																		D						
WP 5	**Outreach**																								
T 5.1	Bringing in new countries																								
T 5.1.1	Value Proposition and KPIs						M						D												
T 5.1.2	Potential Member analysis												D												
T 5.1.3	Facilitating the creation of new national consortia					M						M						M						M	D
T 5.2	Bringing in new users																								
T 5.2.1	Bringing in new categories of users						M						M						M						M
T 5.2.2	Course and educational material registry																		D						
WP 6	**Governance Sustainability**																								
T 6.1	Analysis of governance and management issues																		M				D		
T 6.2	Financial analysis																						D		
T 6.3	Sustainability												M										D		

T 2.1.1 - Extending the SPF coverage

- Responsible: Dieter Van Uytvanck (CLARIN ERIC)
- Goals:
  - include all Identity Providers from all members into the SPF
  - and if feasible all other European federations (CH, HR, ES, FR, GR, HU, IE, LV) and the US federation

T 2.1.2 - Robust technical foundations for the SPF

- Responsible: Jozef Misutka (CUNI)
- Goals:
  - Re-engineer the CLARIN Identity Provider (new front-end, better password hashing) and install it at a computing centre to ensure high availability. This can be used by those without an own Identity Provider or Federation. Conduct a security audit by external security team e.g., CESNET FLAB14.
    - *In cooperation with Willem Elbers and Sander Maijers.*
  - Improve the currently available SAML metadata aggregation workflow.
    - *In cooperation with Sander Maijers.*

T 2.1.2 - Robust technical foundations for the SPF

- Set up monitoring to ensure that there are no missing Identity Providers in the aggregated SAML metadata which is offered. Provide a separate list of malfunctioning Identity Providers for possible feedback. Blacklist clearly non-academic IdPs.
- Collect statistics of attributes released by Identity Providers and use them to improve availability of CLARIN services.
- Failover (redundant) setup of the host that runs infra.clarin.eu and discovery.clarin.eu. Allow Service Providers to include Identity Providers from additional federations in the discovery service.
  - *In cooperation with Willem Elbers, who has already dockerized the discovery service*

T 2.2.1 - Metadata benchmarking and curation

- Responsible: Matej Durco (CLARIN > OEAW)
- Goals:
  - Defining metadata quality score, based on measurable criteria
  - Creating an API (at least Java compatible) to measure quality (score) per CMDI file, this can then be incorporated into a CMDI quality checker (connected to harvester) and to the scoring of the VLO.
  - Development of an application which can automatically analyze large amounts of CMDI metadata records and provide a quality score per record and collection
  - Development of a curation module
  - *In cooperation with Menzo Windhouwer (harvester), VLO developers and the CLARIN-PL colleagues (metadata workflow + harvester)*

T 2.2.2 - Improved harvesting workflow

- Responsible: Marcin Pol (CLARIN > Wrocław)
- Goals:
  - Analyze current issues with the OAI-PMH harvester, e.g.:
    - Lack of robustness
    - Lack of clear error messages for the OAI providers
  - Specify concrete actions to deal with these issues
  - Improve the current (java) code base of the harvester
  - Analyze the triggering of the VLO importer after the harvesting and propose an optimal and robust setup
  - *In cooperation with Menzo Windhouwer (harvester) and Willem Elbers (server setup VLO)*

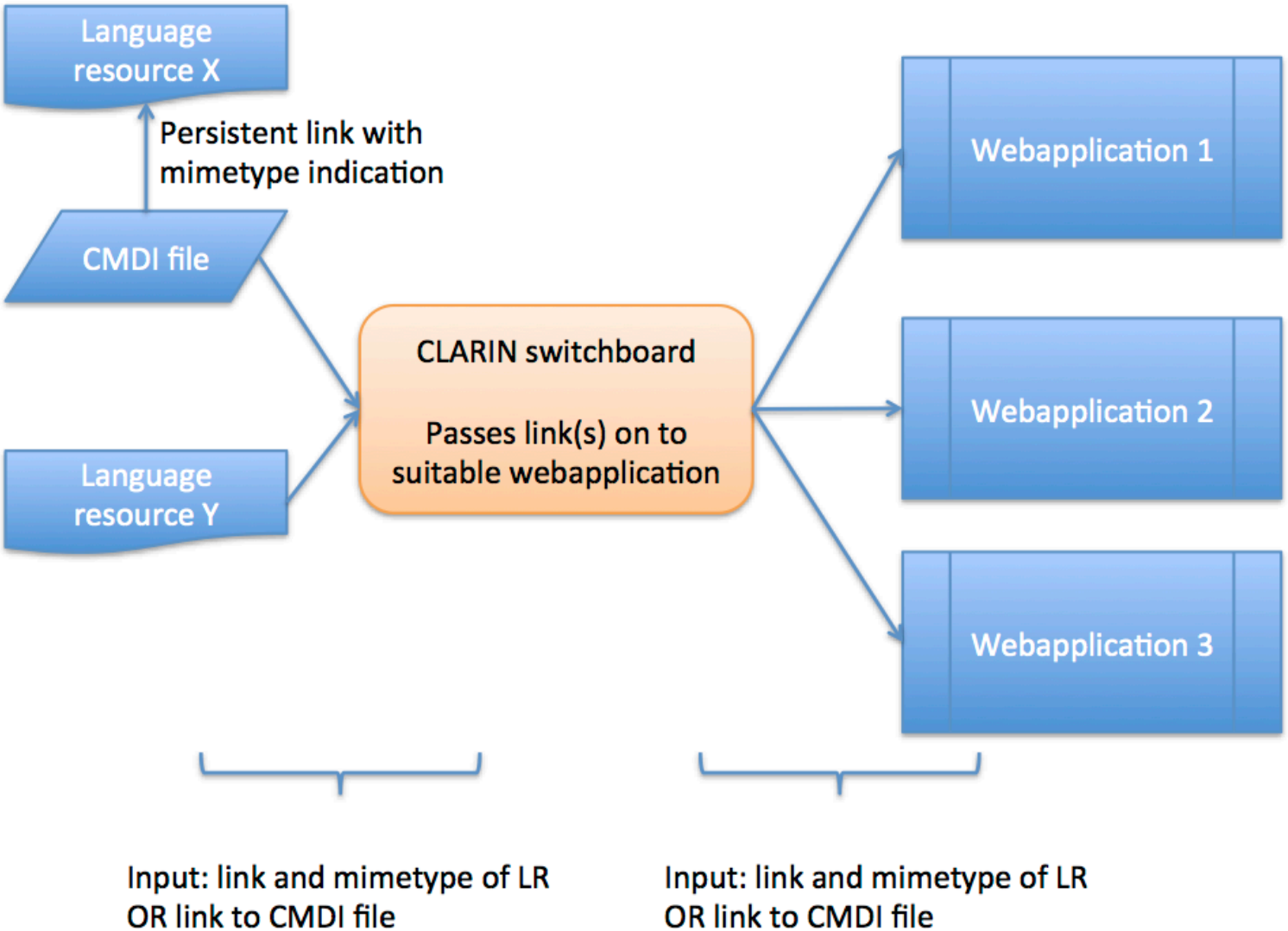
T 2.2.3 - Reinforced concept registry

- Responsible: Menzo Windhouwer (CLARIN > Meertens)
- Goals:
  - Specify requirements for CLARIN use of OpenSKOS, e.g. relations among concepts, easier import of new concepts, concept lifecycle management and browse and search facilities for unauthenticated users.
  - Try to align this with future versions of OpenSKOS or a follow-up platform by the Dutch Cultural Heritage Agency
  - Implement the required features into OpenSKOS
- Integration of OpenSKOS into CMDI 1.2, further development of CMDI 1.2

T 2.3.1 - Language Resource Switchboard

- Responsible: Emanuel Dima (EKUT)
- Goals:
  - Specify in detail the LRS setup
  - Develop a backend service that can match incoming resource metadata with a service
  - Develop a frontend web application for the backend service

T 2.3.1 - Language Resource Switchboard



T 2.3.2 - Integration of new web services into workflows

- Responsible: Emanuel Dima (EKUT)
- Goals:
  - Provide guidance for CLARIN centres to integrate
    - their webservices into workflow engines
    - their web applications into the Language Resource Switchboard
- Provide documentation and tools to make this a smooth process

T 2.3.3 - EUDAT service compatibility

- Responsible: Emanuel Dima (EKUT)
- Goals:
  - Make an analysis of use cases for bringing the code to the data
  - Integrate EUDAT's Generic Execution Framework into CLARIN workflow engines
  - Provide guidance and documentation about providing web services via GEF containers

T 2.4.1 - (Federated) Content Search Engine

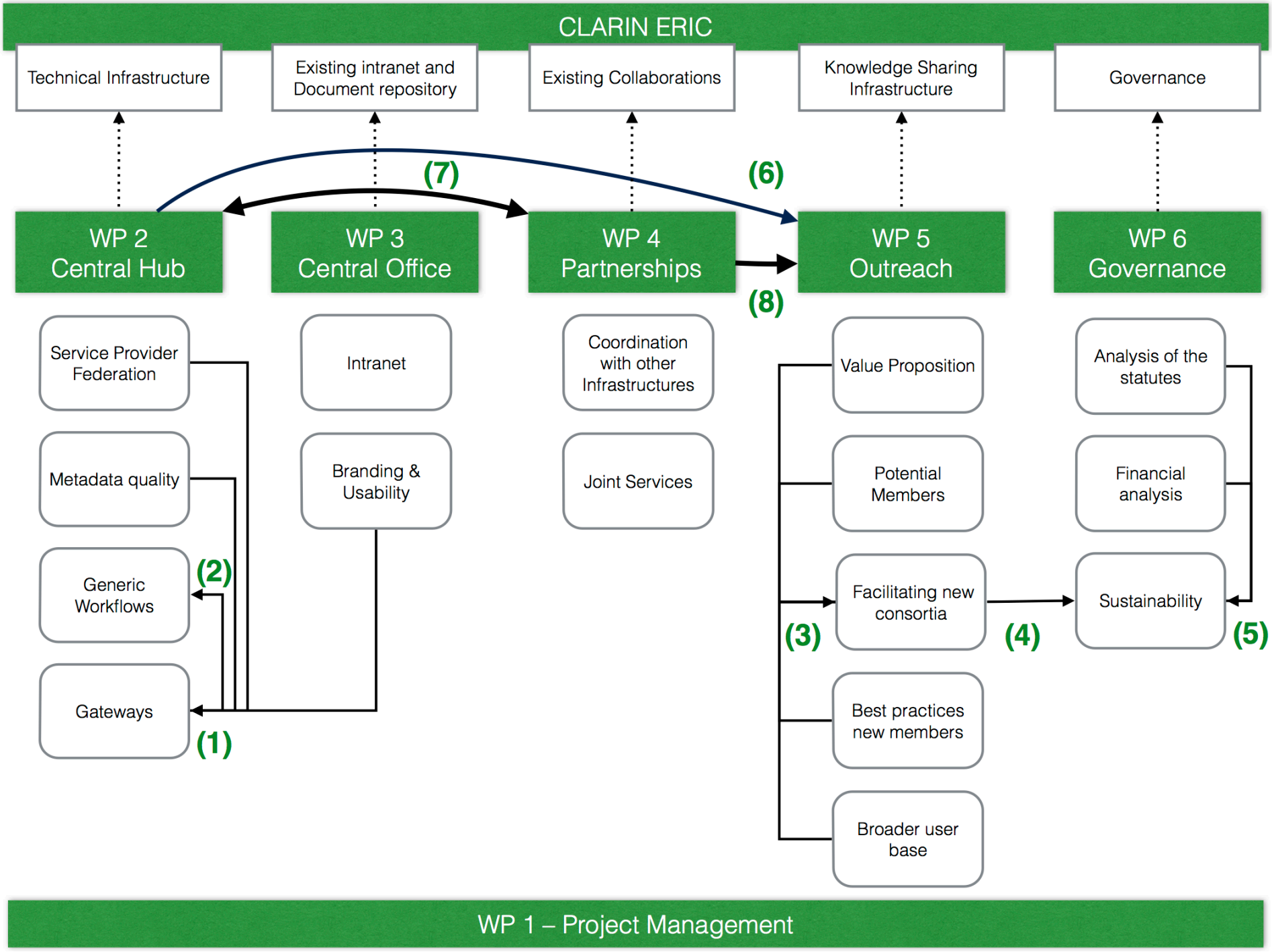
- Responsible: Dieter Van Uytvanck (CLARIN ERIC)
- Work done by: Leif-Jöran Olsson (CLARIN > Språkbanken, 5 PM), Oliver Schonefeld (CLARIN > IDS, 3 PM)
- Goals:
  - Specification of a detailed workplan
  - 2 major iterations, including:
    - Full implementation of FCS 2.0 spec
    - Java library available for endpoints
    - improved icinga plugin to detect time-outs

T 2.4.2 - Virtual Language Observatory

- Responsible: Twan Goosen (CLARIN ERIC)
- Goals:
  - Preparatory work on component registry + CMDI 1.2
  - Specification of a detailed workplan
  - 2 major iterations

T 2.4.3 - Virtual Collection Registry

- Responsible: Willem Elbers (CLARIN ERIC)
- Goals:
  - Preparatory work on component registry + CMDI 1.2
  - Specification of a detailed workplan
  - 2 major iterations


CLARIN ERIC
Technical Infrastructure
Existing intranet and Document repository
Existing Collaborations
Knowledge Sharing Infrastructure
Governance
(7)
(6)
WP 2 Central Hub
WP 3 Central Office
WP 4 Partnerships
WP 5 Outreach
WP 6 Governance
(8)
Service Provider Federation
Metadata quality
Generic Workflows
Gateways
(2)
(1)
Intranet
Branding & Usability
Coordination with other Infrastructures
Joint Services
Value Proposition
Potential Members
Facilitating new consortia
Best practices new members
Broader user base
(3)
(4)
Analysis of the statutes
Financial analysis
Sustainability
(5)
WP 1 – Project Management

Conclusions

- Many tasks ahead of us …
- … but much of the ground work has been done
- And there is an excellent team to work on it!