ALIGNED
Software and Data Engineering

Engaging Content
Engaging People

Part 4: Dacura/RDF Rollout Plans

Rob Brennan

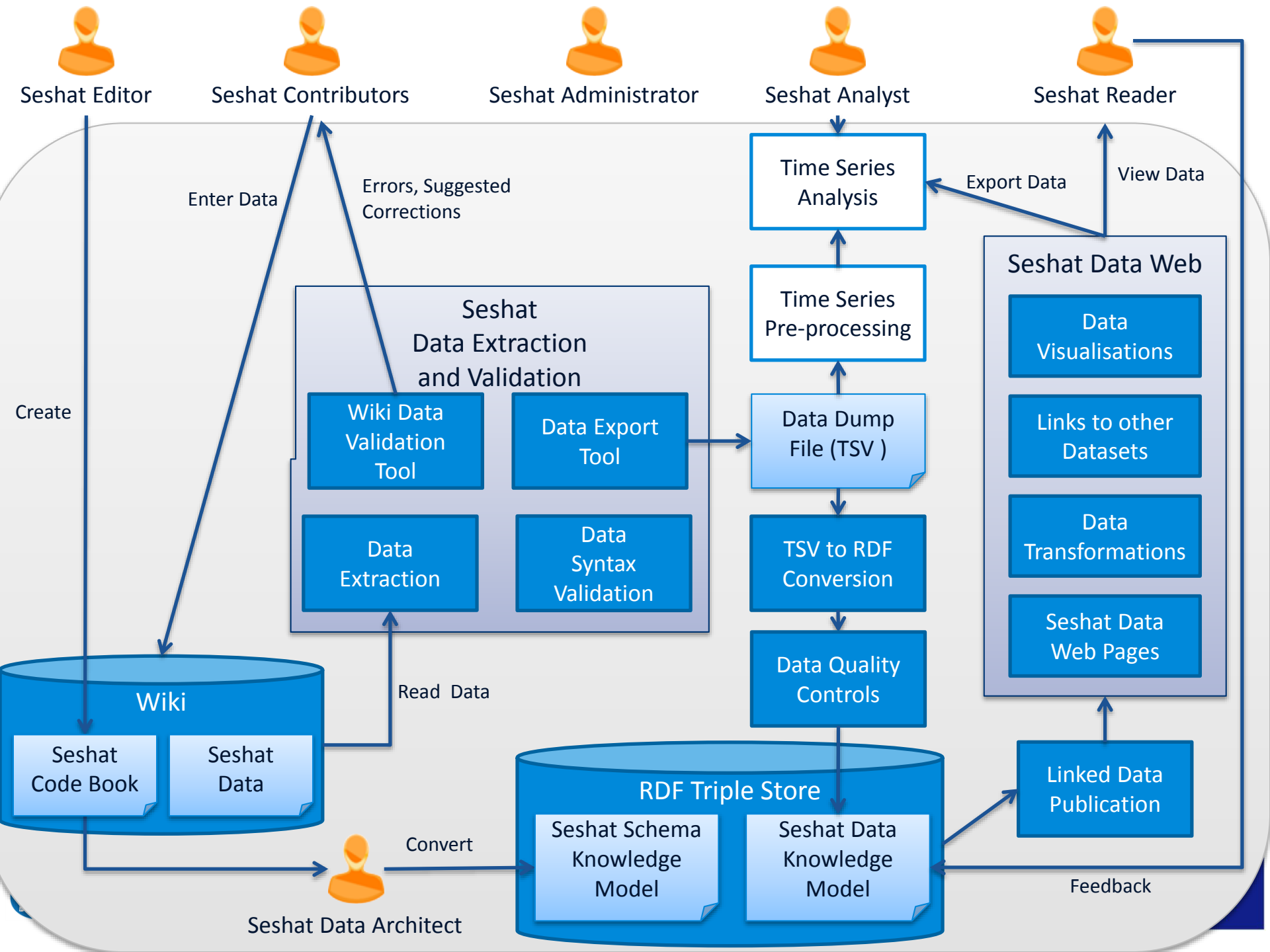
SFI Seshat Workshop May 9th 2015

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.

Agenda

- Reminder: where we are
- October 2015 pilot
  - Proposed feature list
  - Feedback
- September 2016 roll-out
  - Proposed feature list
  - Feedback
- Discussion (post 2016?)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.


Seshat Editor
Seshat Contributors
Seshat Administrator
Seshat Analyst
Seshat Reader
Enter Data
Errors, Suggested Corrections
Create
Time Series Analysis
Export Data
View Data
Seshat Data Web
Data Visualisations
Links to other Datasets
Data Transformations
Seshat Data Web Pages
Time Series Pre-processing
Seshat Data Extraction and Validation
Wiki Data Validation Tool
Data Export Tool
Data Extraction
Data Syntax Validation
Data Dump File (TSV)
TSV to RDF Conversion
Data Quality Controls
Wiki
Seshat Code Book
Seshat Data
Read Data
RDF Triple Store
Seshat Schema Knowledge Model
Seshat Data Knowledge Model
Linked Data Publication
Convert
Seshat Data Architect
Feedback

Two Ways to Look at Scope

- Dacura feature list
  - Focus on system components/development tasks
  - Computer science/IT focus
  - Becomes more fine-grained over time

- Seshat Use Cases
  - "What will the system do from a user point of view?"
  - In ALIGNED Deliverable D2.1… this will have 3 versions during the project (v1 May 2015)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.

October 2015 pilot overview

ALIGNED
Software and Data Engineering

- Goals:
  - Live RDF snapshot of Seshat data
  - Bring Oxford CS into development
  - Enable feedback from user trials in Oxford/AMU Poznan
  - Create baseline system to evolve into Aug 2016 launch
- Caveats
  - It cannot be used for “real work”
  - Most functionality is platform development, not cool stuff
  - No assurances about bugs/usability/stability
  - The Seshat ontology **will change** for the released version

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

UNIVERSITY OF OXFORD

UNIVERSITÄT LEIPZIG

SEMANTIC WEB COMPANY
linking data to knowledge

Wolters Kluwer

UAM

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 644055.

Oct 2015 Feature List

ALIGNED
Software and Data Engineering

- Seshat full pipeline v1
  - Wiki editing data capture widget, quality control/validation, publish to RDF, push back to wiki
- Candidate API
- Basic publishing to wiki, TSV datadumps, HTML, linked data
- Candidate generation from wiki
- Schema versioning support and change management
- Schema best practice quality service

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

UNIVERSITY OF OXFORD

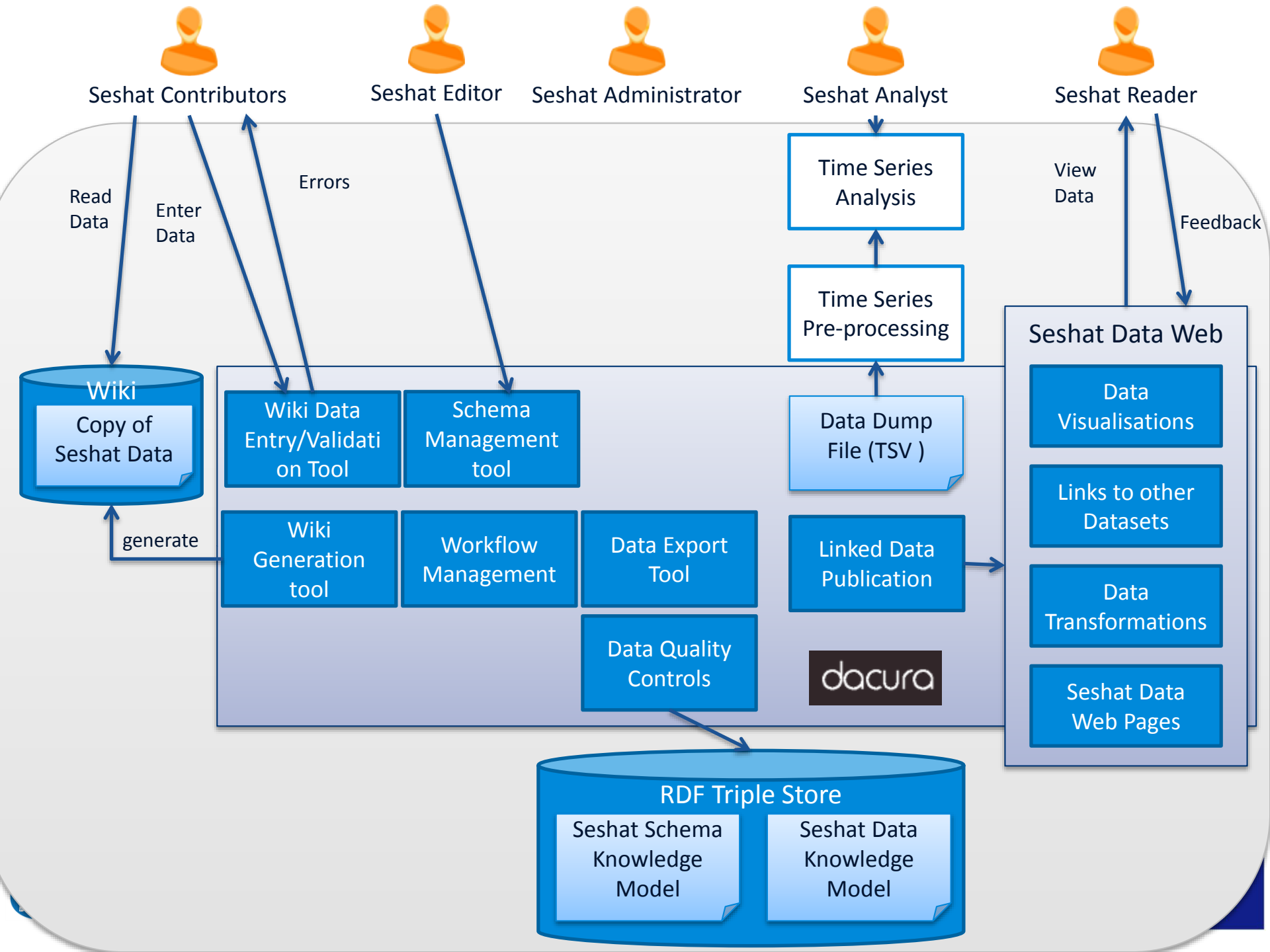
UNIVERSITÄT LEIPZIG

SEMANTIC WEB COMPANY
linking data to knowledge

Wolters Kluwer

UAM

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.


Seshat Contributors
Seshat Editor
Seshat Administrator
Seshat Analyst
Seshat Reader
Read Data
Enter Data
Errors
View Data
Feedback
Time Series Analysis
Time Series Pre-processing
Seshat Data Web
Wiki
Copy of Seshat Data
Wiki Data Entry/Validation Tool
Schema Management tool
Data Dump File (TSV)
Data Visualisations
Links to other Datasets
Data Transformations
Seshat Data Web Pages
generate
Wiki Generation tool
Workflow Management
Data Export Tool
Linked Data Publication
Data Quality Controls
dacura
RDF Triple Store
Seshat Schema Knowledge Model
Seshat Data Knowledge Model

Oct 2015 Use Case 0

- **RDF-based Seshat Data** – the existing Seshat dataset and schema (codebook) shall be available as RDF based on the Seshat Ontology/Data Model.
- **For 2015**:
  - Full version of Seshat Ontology available
  - Tools to convert Seshat wiki to RDF
  - NB Wiki is still canonical data source

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.

Oct 2015 Use Case 1

- **SS1. Data Validation** – a harvester enters a value or set of values into the dataset which is syntactically, semantically or factually invalid. Currently there is very limited support for validation of these values and it is thus easy for incorrect or invalid entries to be added to the dataset.
- **For 2015**:
  - User-interface elements will be generated from the schema and will help users enter only correct data.
  - Dacura can import the current data from the Wiki into an RDF triple-store and validate it.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.

Oct 2015 Use Case 2

- **SS2. Capture of Data Complexity** – the Seshat architects desire that the harvesters express the full complexity of the data – and capture where values are uncertain or disputed. However, the harvesters tend to prioritise speed over complexity for a variety of reasons and will often neglect to express the full complexity of the evidence. Another problem is that, when dealing with pre-historical societies, the data required by the schema requires significant interpretation which is often beyond the competence or confidence of data-harvesters.
- **For 2015**:
  - Seshat Ontology captures more structured data than the wiki.
  - User-interface elements will expose more complex structures for data provenance, temporal scoping.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.

Oct 2015 Use Case 3

- **SS3. Schema Evolution –** the Seshat schema has been developed iteratively and continues to evolve. Data that has been collected with earlier versions of the schema currently needs to be manually updated to make it consistent with schema updates.
- **For 2015**:
  - Dacura schema versioning support and change management will assess and report the impact of proposed schema changes.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.

Oct 2015 Use Case 4

- **SS4. Dataset Evolution –** the Seshat dataset has been in rapid evolution since its inception and is expanding at an increasing rate. The Seshat researchers would like to continue to increase the rate at which high-quality data is added to the system. They would also like to gain greater understanding of how the dataset has evolved: in what context was a given variable added? Why was a value changed? How do the overall characteristics of the dataset change over time?
- **For 2015**:
  - Dacura schema versioning support and change management will automatically add structured provenance information to all dataset changes.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.

Oct 2015 Use Case 5

ALIGNED
Software and Data Engineering

- **RA-based Data Collection –** the current wiki-based Seshat data collection process will be supported but based on the RDF Seshat data model.
- **For 2015**:
  - Continue to use seshat wiki for human navigation/display of data
  - Replace wiki editing and Dacura data validation tool with Dacura full Seshat pipeline version 1:
    Wiki editing data capture widget -> data quality control/validation -> publish to RDF -> push back to wiki

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

UNIVERSITY OF OXFORD

UNIVERSITÄT LEIPZIG

SEMANTIC WEB COMPANY
linking data to knowledge

Wolters Kluwer

UAM

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.

Oct 2015 Use Case 6

- **Multi-format Data Publication –** the Seshat data needs to be consumed by four target Seshat roles: editors, contributors, knowledge engineers and data analysts.
- **For 2015**:
  - Continue to use seshat wiki for human navigation/display of data.
  - the Seshat data will be made available in RDF as Linked Data in a closed website
  - Dacura Seshat Data Export tool will dump the RDF as a TSV for data analysts

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.

ALIGNED
Software and Data Engineering

2016 Release – Live RDF Data

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

UNIVERSITY OF OXFORD

UNIVERSITÄT LEIPZIG

SEMANTIC WEB COMPANY
linking data to knowledge

Wolters Kluwer

UAM

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.

Aug 2016 Release Feature List

ALIGNED
Software and Data Engineering

- Candidate filter tool
- Training tool
- Productivity management tool
- User access control on sub-datasets
- User management
- DBpedia import/candidate generation
- Live Seshat data as RDF

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

UNIVERSITY OF OXFORD

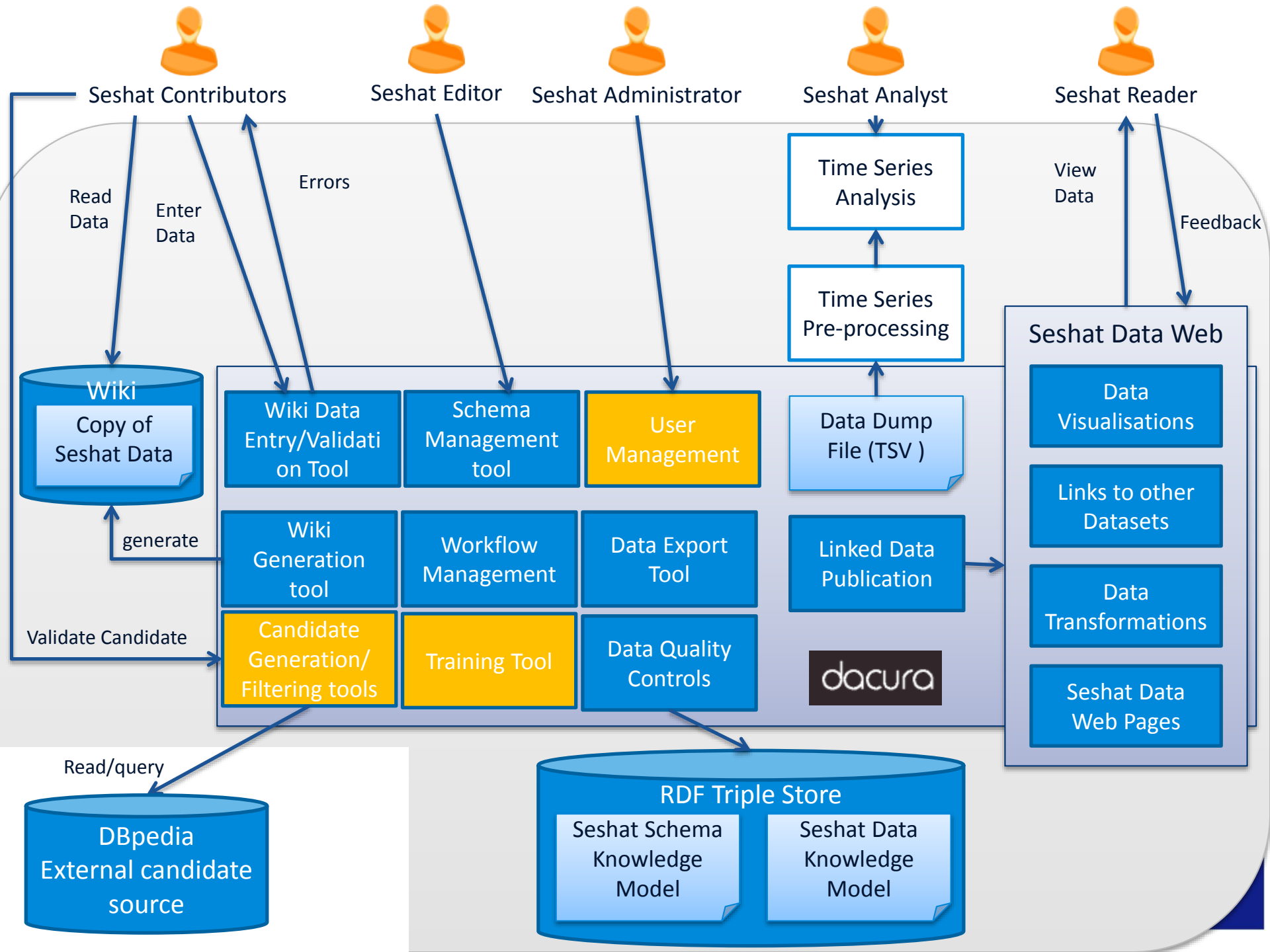
UNIVERSITÄT LEIPZIG

SEMANTIC WEB COMPANY
linking data to knowledge

Wolters Kluwer

UAM

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.


Seshat Contributors
Seshat Editor
Seshat Administrator
Seshat Analyst
Seshat Reader
Read Data
Enter Data
Errors
View Data
Feedback
Time Series Analysis
Time Series Pre-processing
Seshat Data Web
Wiki
Copy of Seshat Data
Wiki Data Entry/Validation Tool
Schema Management tool
User Management
Data Dump File (TSV)
Data Visualisations
Links to other Datasets
Data Transformations
Seshat Data Web Pages
generate
Wiki Generation tool
Workflow Management
Data Export Tool
Linked Data Publication
Validate Candidate
Candidate Generation/ Filtering tools
Training Tool
Data Quality Controls
dacura
Read/query
DBpedia External candidate source
RDF Triple Store
Seshat Schema Knowledge Model
Seshat Data Knowledge Model

Aug 2016 Updated Use Cases (i)

- **Use Case 0: RDF-based Seshat Data**.
- **For 2016**:
  - RDF version of Seshat is now canonical
  - Wiki is for human navigation/publication only
- **Use Case 1. Data Validation** –.
- **For 2016**:
  - Improved usability data input interfaces based on 2015 trials.
  - Data Input training tool for RAs to increase productivity/reduce errors/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.

2016 Updated Use Cases (ii)

- **Use Case 2. Capture of Data Complexity** .
- **For 2016**:
  - Improved Seshat Ontology based on 2015 trials.
  - Improved user-interface elements based on 2015 trials.
- **Use Case 3. Schema Evolution –**.
- **For 2016**:
  - No changes.
- **Use Case 4. Dataset Evolution –**
- **For 2016**:
  - No changes.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.

2016 Updated Use Cases (iii)

ALIGNED
Software and Data Engineering

- **Use Case 5: RA-based Data Collection**
- **For 2016**:
  - Improved usability data input interfaces based on 2015 trials.
- **Use Case 6: Multi-format Data Publication**.
- **For 2015**:
  - No changes

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin
UNIVERSITY OF OXFORD
UNIVERSITÄT LEIPZIG
SEMANTIC WEB COMPANY
linking data to knowledge
Wolters Kluwer
UAM

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.

Oct 2016 Use Case 7

ALIGNED
Software and Data Engineering

- **Dbpedia-based Candidate Generation –** need to automate generation of candidates. DBpedia will be first target.
- **For 2016**:
  - DBpedia import of candidates
  - Candidate filtering tool for RAs
  - Candidate source management tool

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

UNIVERSITY OF OXFORD

UNIVERSITÄT LEIPZIG

SEMANTIC WEB COMPANY
linking data to knowledge

Wolters Kluwer

UAM

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.

Oct 2016 Use Case 8

ALIGNED
Software and Data Engineering

- **User management and dataset productivity –** need to empower Seshat administrators to manage users, limit the scope of their work and automate generation of candidates. Dataset managers need to monitor the progress of dataset collection.
- **For 2016**:
  - Productivity management tool
  - User access control on sub-datasets
  - User management

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.

ALIGNED
Software and Data Engineering

Dacura Beyond 2016

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

UNIVERSITY OF OXFORD

UNIVERSITÄT LEIPZIG

SEMANTIC WEB COMPANY
linking data to knowledge

Wolters Kluwer

UAM

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.

From ALIGNED Workplan

- Data Federation
  - Model mapping tool – Linked Data source management
  - Data provenance IP tracking
- Domain Expert interpretation tool
- Advanced User management
  - Work allocation and scheduling
- Access controlled publishing
- Data visualisation tools
- Hosting 3rd party datasets

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.

ALIGNED
Software and Data Engineering

ADAPT
Engaging Content
Engaging People

Session 5
- Codebook/Seshat Update Processes

Rob Brennan

SFI Seshat Workshop May 9th 2015

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.

Current codebook update process

- Codebook specified on wiki
- Seshat editorial board oversees changes to codebook
- Changes occur on an “as needed” basis
- However in practice codebook changes when:
  - New expert communities join
  - New projects address new “big questions”
  - RAs provide feedback
  - People have ideas about what might be interesting

Discussion: what have we missed?

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 644055.

Issues with current process

- Wiki pages remain in old codebook format
  - RA effort harnessed to update (sometimes)
  - As size of dataset increases the impact grows (exponentially?)
  - Relevant experts sometimes no longer active
  - Dacura Seshat scraper code sometimes needs to be updated for structural changes in wiki
- Informal email discussions are hard to follow/track
- No formal assessment of likely impact of changes
- Unclear when changes are rolled out

Discussion: What other issues do people have?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.

Technology implications for future processes - internal

- RDF is not as flexible as the wiki
- Codebook changes will have to be expressed in Seshat Ontology
  - Needs input from knowledge engineers
- Dacura schema management tool will allow us to evaluate for a given codebook change:
  - How many variable instances will be impacted
  - How will this impact the consistency and integrity of the current seshat dataset?
  - Degree to which updates can be automated
  - (Maybe) estimate of RA work and Expert work in person-hours to implement the change

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.

Technology implications for future processes - external

- If there are tools built on top of Seshat data eg visualisations then they could be impacted too
  - Worst case: Dacura tools are impacted and need a new release
- Federated data models could break
- Data consumers of seshat data will be impacted
- If we publish external schema then mappings could break
- Imported data-sets/candidates could need to be re-run
- Dacura mapping management tool (2017?) could address some of these issues

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.

Discussion on future update processes

- Strawman Proposal:
- Formal 6 monthly codebook meeting at Seshat workshops
- Establish a set of policy guidelines for evaluating codebook changes
- Dacura schema management tool to be used to assess the impact of codebook changes.
- Establish a process for roll-out of codebook changes

Discussion 1: How practical is this strawman?

Discussion 2: What are these policies/processes?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.

- Done!

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin
UNIVERSITY OF OXFORD
UNIVERSITÄT LEIPZIG
SEMANTIC WEB COMPANY
linking data to knowledge
Wolters Kluwer
UAM

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644055.