The CRESCAT Platform

A Computational Research Environment for Scientific Collaboration on Ancient Topics

PI: David Schloen, University of Chicago

Goals

- **Connect existing software** to support multi-project and multi-disciplinary research
- **Support all stages of data:**

Acquisition → Integration → Analysis → Publication → Archiving

- **Accommodate heterogeneity** of data sources, data types, and logical schemas while preserving original ontologies/terminologies
- **Automate data transformations** and transfers from one stage to the next via high-level GUI
- **Provide seamless scalability** for large-scale data management and various algorithmic analyses
- **Ensure sustainability** of:
  - – Software maintenance
  - – Technical support
  - – Institutional hosting
- **Test and document the platform** with complex use cases from archaeology & ancient economics

Acknowledgments

This work is supported by NSF SI2-SSI via award no. ACI-1450455.

1. Data Acquisition

- **External database APIs**
  - – Live links to curated data stored in external repositories
  - – Granularity (atomization) to the extent permitted by the API
- **Mobile apps for fieldwork**
  - – Offline mode with automated syncing when online
  - – Accommodates wide variability in project schemas
- **Instrumental data**
  - – Supports many data types and file formats (image, audio, video, GIS, 3D models, etc.)

2. Data Integration

- **Ontology-agnostic**
  - – Represent explicitly both research data and the disparate ontologies inherent in the data
  - – Abstract top-level ontology implemented in a non-relational schema (keyed and indexed)
- **Data warehouse approach**
  - – XQuery DBMS optimized for hierarchical semi-structured data
  - – Atomic keyed data objects correspond to project-defined units of *space, time*, *agency*, *taxonomy,* etc.
  - – Automatic parsing of source data to populate the data warehouse

3. Data Analysis

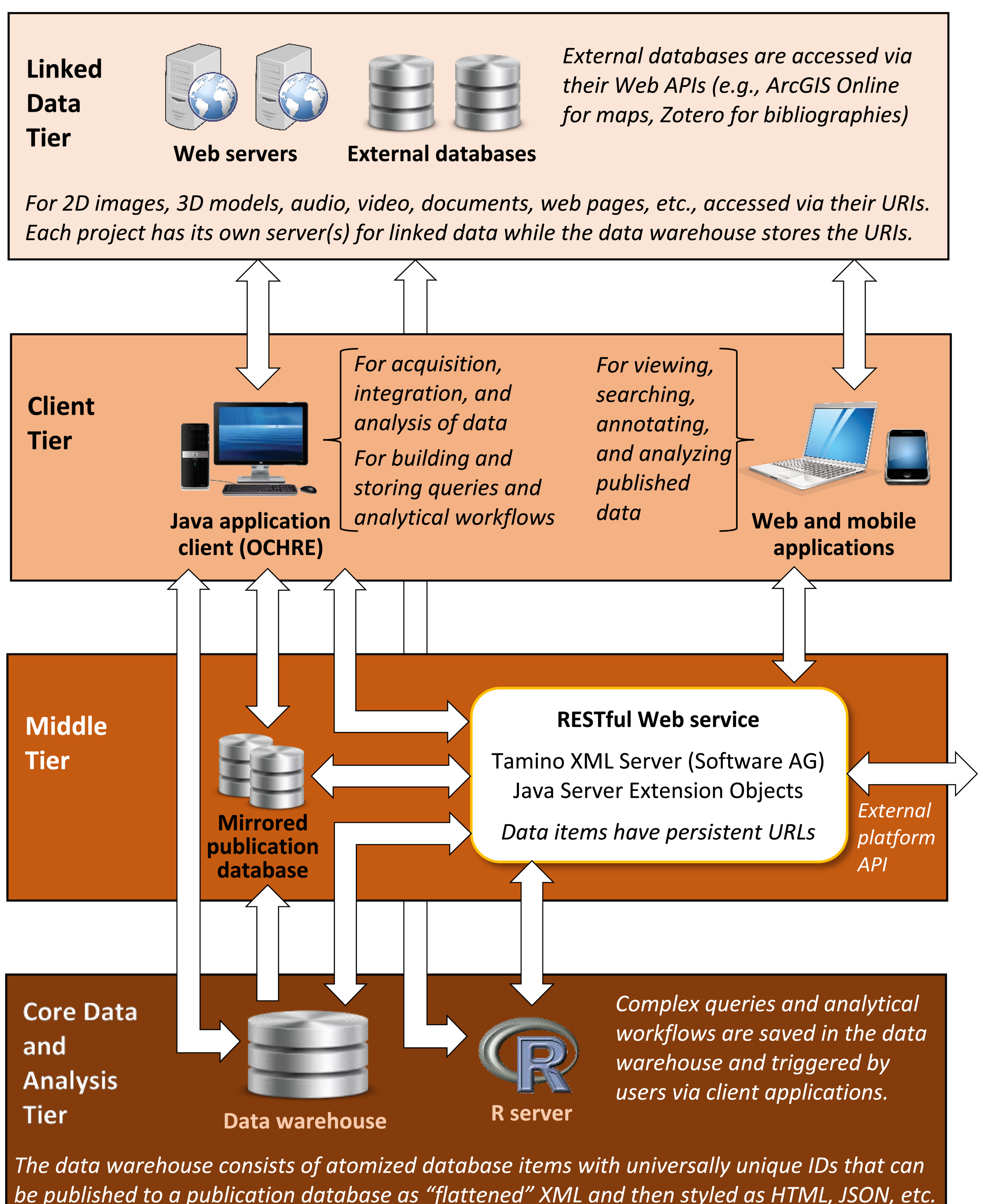
- **Complex queries**
  - – Leverage overlapping hierarchical taxonomies with inheritance
  - – Spatial and temporal relations implied by hierarchies and links
- **Statistical analysis and viz**
  - – Close integration with scalable R server using workflow scripts
- **Geospatial analysis**
  - – Close integration with ArcGIS
  - – Object-based ownership of vector/raster data better than typical layer-based model
- **Social network analysis**

4. Data Publication

- **REST API for Web apps**
  - – Flattened views of XML data selected via queries, with XSLT stylesheets to render as JSON

5. Data Archiving

- **OWL ontology specification**
  - – Defines top-level ontology of the data warehouse
- **RDF triples reflect ontology**
  - – Export data from a given research project as RDF triples that preserve all atomized distinctions and relationships



Sustainability

- **Nonproprietary standards**
  - – Based entirely on XML, RDF, and related W3C standards
- **Proven business model**
  - – Institutional support of computing infrastructure and modest fees to cover technical support

Dissemination

- **Advisory board members** from several universities representing several different disciplines
- **Workshops and conferences** to get user feedback
- **Website with documentation** of software and methods

Example Use Case

- **Ancient Greek economy and trade**
  - – Thousands of coin hoards have been found around the Mediterranean Sea, dating to the earliest use of coins (600–300 BC)
  - – Network analysis of spatial and temporal distributions of coins by type, date, mint of origin, and findspot lets us infer economic and political structure