CRESCAT
A Computational Research Environment for Scientific Collaboration on Ancient Topics

PI: David Schloen, University of Chicago

Goals

- **Support all 5 stages of data for** multi-disciplinary collaborative research
- **Automate data transfers and transformations** from one stage to the next via high-level GUI
- **Accommodate hetero-geneity** of data sources, types, and schemas while preserving the original ontologies
- **Seamless scalability** for data management and algorithmic analyses
- **Ensure sustainability** of software maintenance and technical support
- **Test and document** with complex use cases from
  - ❖ Archaeology
  - ❖ Paleontology
  - ❖ Historical linguistics
  - ❖ Ancient economics
  - ❖ Population genetics
  - ❖ Paleocliimatology
  - *etc.*

1. Acquisition

- **External curated data repositories** via live links using their Web APIs
- **Instruments and data files** with support for many data types and file formats (2D images, 3D models, audio, video, geospatial, etc.)
- **Manual entry** with offline mode for field input and automated syncing of data when back online

2. Integration

- **Ontology-agnostic data warehouse** stores both data and the ontologies inherent in the data
- **XQuery DBMS** optimized for hierarchies of atomic keyed data objects representing spatial, temporal, linguistic, and taxonomic relationships
- **Automatic parsing** of source data to populate the integrated warehouse

3. Analysis

- **Complex queries** use hierarchical taxonomies with semantic inheritance
- **Statistical analysis and visualization** via tightly integrated R server with data-aware console
- **Geospatial mapping and analysis** via ArcGIS Online and ESRI components

Example Use Case

Ancient Greek economy via network analysis of thousands of coin hoards

SOCIAL NETWORK ANALYSIS John Scott

4. Publication

- **REST API** exposes published data as XML with XSLT stylesheets to render it as JSON/HTML
- **Sample Web apps** provided for various research domains, to be customized as needed
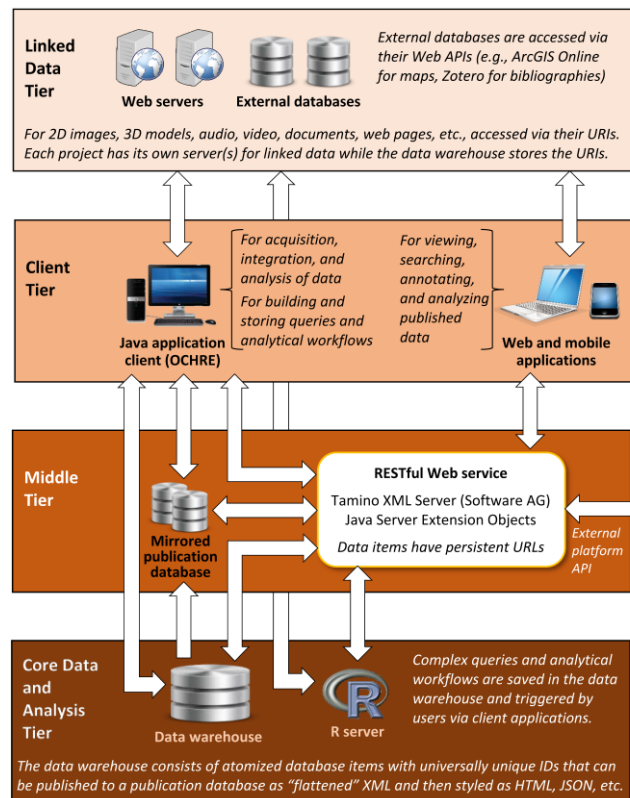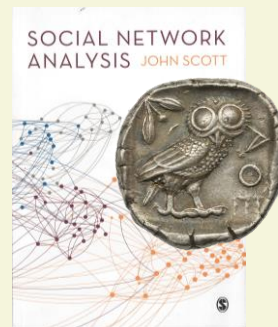- **Secure, password-protected data** controlled and published by owners

5. Archiving

- **OWL-RDF ontology specification** documents the top-level (upper) ontology underlying the data warehouse
- **Can export RDF triples** conformant to the OWL ontology, preserving all distinctions and relation-ships in the data, for use in other graph databases

Linked Data Tier

Web servers   External databases

External databases are accessed via their Web APIs (e.g., ArcGIS Online for maps, Zotero for bibliographies)

For 2D images, 3D models, audio, video, documents, web pages, etc., accessed via their URIs. Each project has its own server(s) for linked data while the data warehouse stores the URIs.

Client Tier

Java application client (OCHRE)

For acquisition, integration, and analysis of data
For building and storing queries and analytical workflows

For viewing, searching, annotating, and analyzing published data

Web and mobile applications

Middle Tier

Mirrored publication database

RESTful Web service
Tamino XML Server (Software AG)
Java Server Extension Objects
Data items have persistent URLs

External platform API

Core Data and Analysis Tier

Data warehouse   R server

Complex queries and analytical workflows are saved in the data warehouse and triggered by users via client applications.

The data warehouse consists of atomized database items with universally unique IDs that can be published to a publication database as "flattened" XML and then styled as HTML, JSON, etc.