Applying 21st Century Technology to the World's Longest-Lived Oceanographic Data Set

Ardys Kozbial, University of California San Diego Libraries

PRDLA, October 18, 2007

Disclaimer

This project is filled with possibilities.

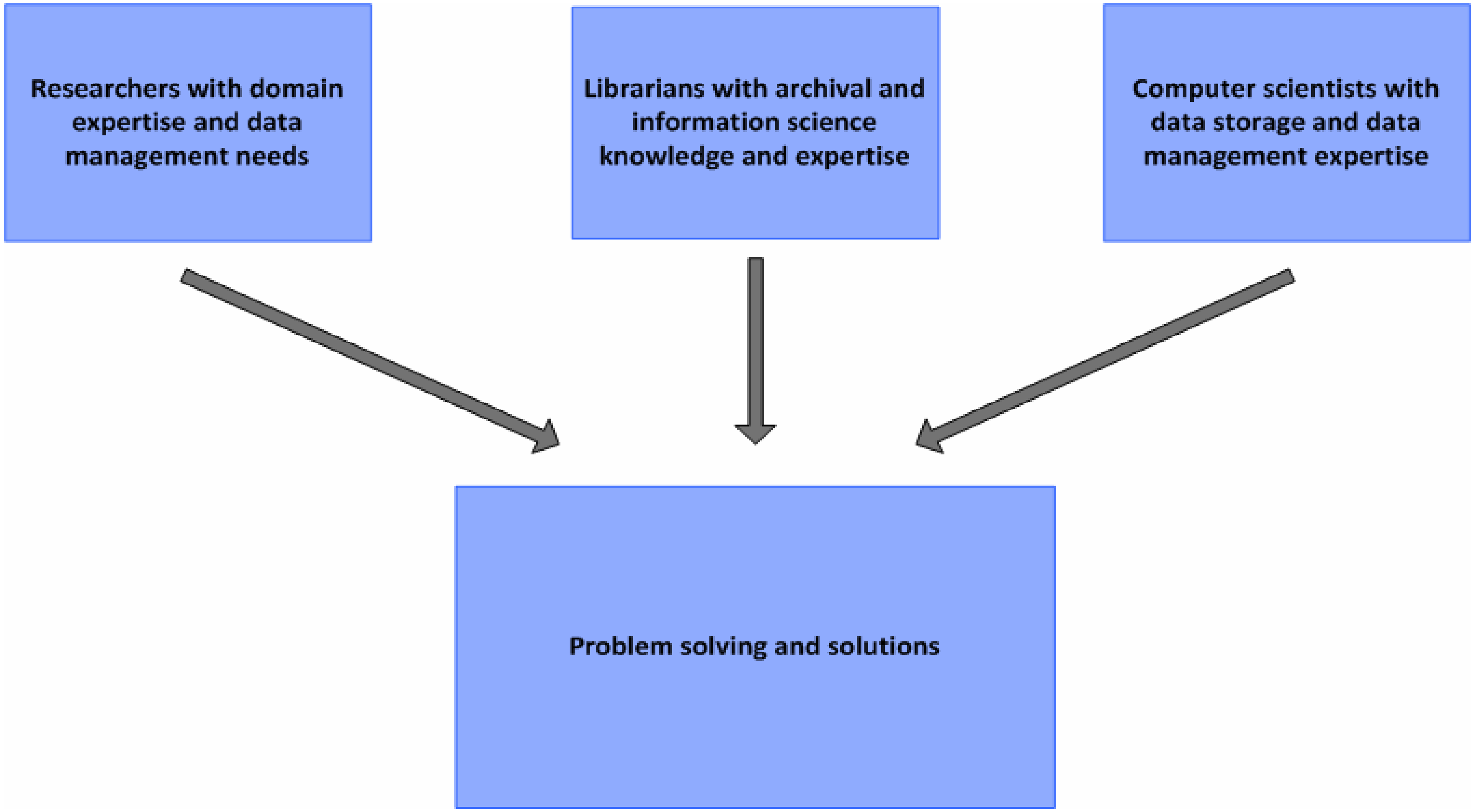
Collaboration

- Partners
  - Scripps Institution of Oceanography (SIO)
  - San Diego Supercomputer Center (SDSC)
  - UCSD Libraries

Collaborative Projects to Date

- Digital Preservation (infrastructure)
  - DAMS, Chronopolis
- Collection Ingest
  - UCTV videos, LC image collection pilot, web archives
- Digital Preservation (content)
  - Multi-Institution Testbed for Scalable Digital Archiving (DigArch)

Expertise



Researchers with domain expertise and data management needs

Librarians with archival and information science knowledge and expertise

Computer scientists with data storage and data management expertise

Problem solving and solutions

Competencies Leveraged

Faculty

□ Domain expertise
□ Data collection
□ Taxonomies
□ Ontologies
□ Data mining
□ Data reuse

Libraries

□ Archiving
□ Metadata management
□ Discovery-tool building
□ Culture of service
□ Culture of trust
□ Project Management

SDSC

□ Grid storage
□ Grid services
□ Data management
□ Data preservation
□ Format migration

Find Data that Need Care and Feeding

- California Cooperative Oceanic Fisheries Investigations
  - CalCOFI
    - The California Cooperative Oceanic Fisheries Investigations (CalCOFI) are a unique partnership of the California Department of Fish and Game, the NOAA Fisheries Service and the Scripps Institution of Oceanography. The organization was formed in 1949 to study the ecological aspects of the collapse of the sardine populations off California. Today its focus has shifted to the study of the marine environment off the coast of California and the management of its living resources. The organization hosts an annual conference, publishes data reports and a scientific journal and maintains a publicly accessible data server (www.calcofi.org).

Map of the CalCOFI Stations

- http://www.calcofi.org/newhome/cruises/station_map.htm

Larger Map

http://maps.google.com/

How is use of the data changing?

- Originally marine biologists studying fish populations
- Now
  - Marine biologists
  - Climatologists
  - Bioinformatics
  - Zoologists
- Data reuse

What are the possibilities?

- Visualization toolkit/toolbox
- Integrate satellite data with CalCOFI data
  - the GIS project
- Integrate CalCOFI data with Baja California data and Monterrey data
- Cross-repository coordination with keywords, metadata

What is Data Central?

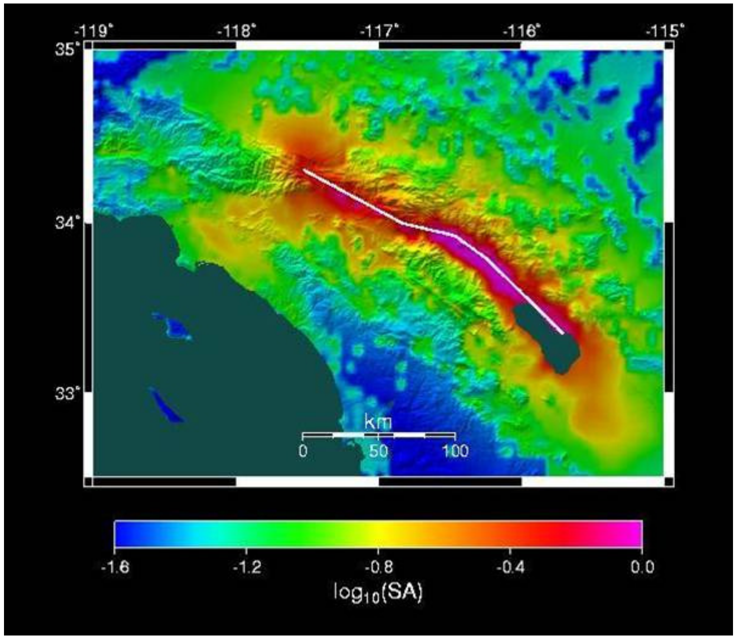
- Comprehensive data environment that incorporates access to the full spectrum of data enabling resources
- First program of its kind to support research and community data collections and databases

Why SDSC Data Central?

- SDSC has experienced increasing demand by the domain communities for collaborations on data driven discovery including
  - hosting, managing, publishing data in digital libraries
  - sharing data through the web and data grids
  - creating, optimizing, porting large scale databases
  - data intensive computing with high bandwidth data movement
  - analyzing, visualizing, rendering and data mining large scale data
  - preservation of data in persistent archives
  - building collections, portals, ontologies
  - providing resources, services, expertise

TeraShake

- TeraShake simulates a 7.7 earthquake along the southern San Andreas fault close to LA using seismic, geophysical, and other data from the Southern California Earthquake Center

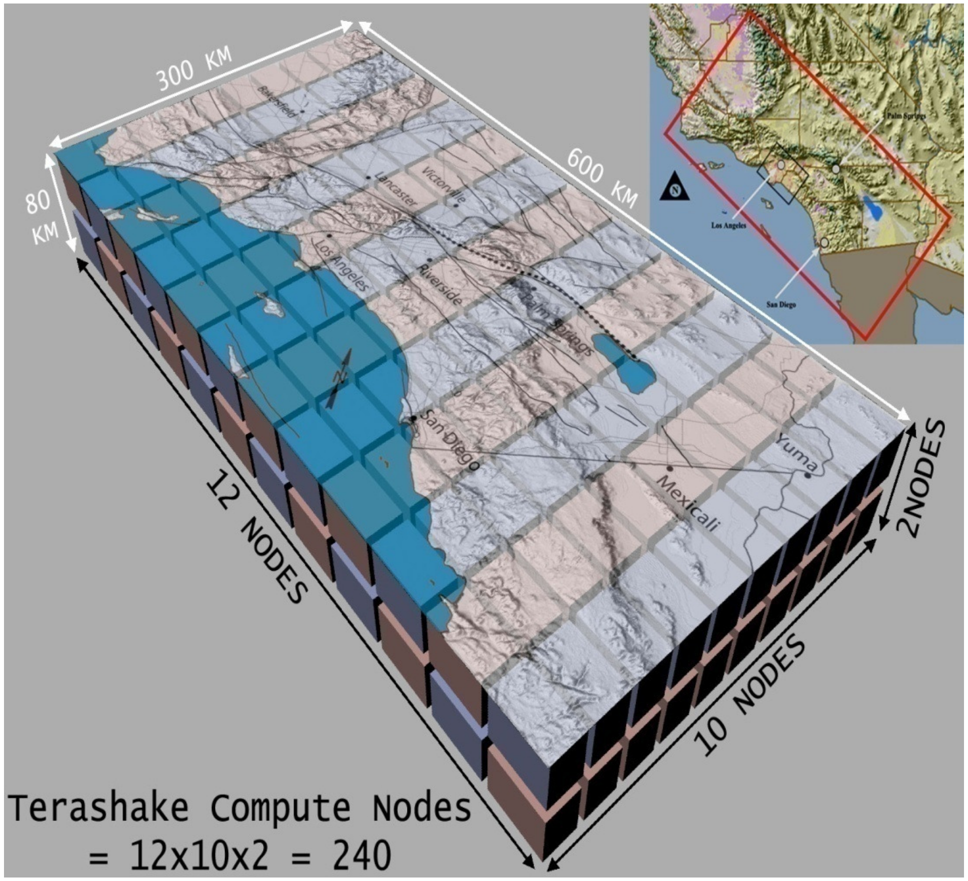


How **TeraShake** *Works*

How TeraShake simulates earthquakes:

1. Divide up Southern California into “blocks”

2. For each block, get all the data on ground surface composition, geological structures, fault information, etc.

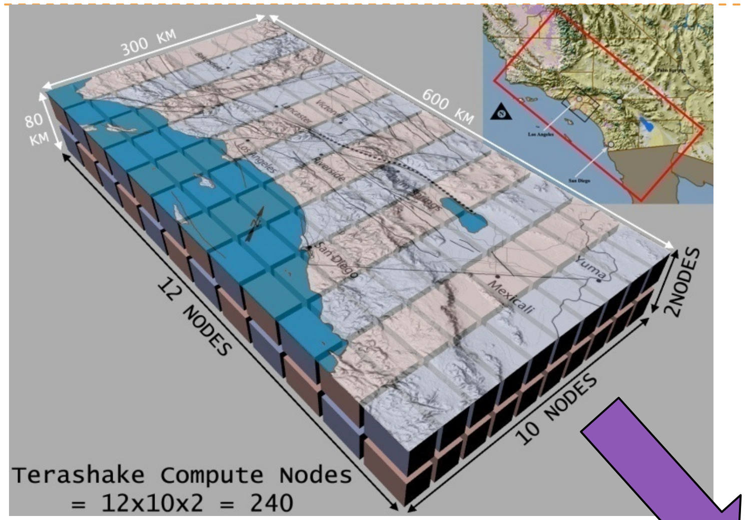


Terashake Compute Nodes
= 12x10x2 = 240

How TeraShake Works

3. Map the blocks on to processors of the supercomputer

4. Run the simulation using current information on fault activity and the physics of earthquakes



SDSC's DataStar – one of the 50 fastest computers in the world

What Libraries Bring to the Table

- Significant expertise
  - Metadata
  - Archival management
  - Policy development
- Organizational experience and stability
  - Process and results driven
- Culture of trust
  - Responsible guardians of the cultural record
  - Service oriented
  - Respectful of privacy and intellectual property

What Libraries Bring to the Table (another view)

- Data acquisition, ingest layer
  - Selection, taxonomy, ontology, metadata, workflow
- Preservation layer
  - Archival retention, format migration, quality assurance, trust
- Physical layer
  - Storage, network security, reliability standards
- Service layer
  - Discovery, retrieval, data mining, data visualization
- Management layer
  - Administration, budget, policy, development

What are the next steps for this project?

- Further discussion
- Focus on one of the possibilities
- Analyze the data and the metadata
- Search for funding opportunities
- Build the tool

Questions?

UCSD Geisel Library and Warren Mall
Courtesy UCSD Publications
Copyright © 1996 by UC Regents

DP_LIBG002-E