



# Genome Data Directories

---

Don Gilbert, [gilbertd@indiana.edu](mailto:gilbertd@indiana.edu)  
May 2003





# Data Access Needs

---

Computable genome data access

- Page scraping and bulk files not enough
- Internet search & retrieval of all genome objects distributed among many sources
- Simple, flexible client program model
- Efficient for high volumes ( $10^5$  objects; >1 GB sizes)





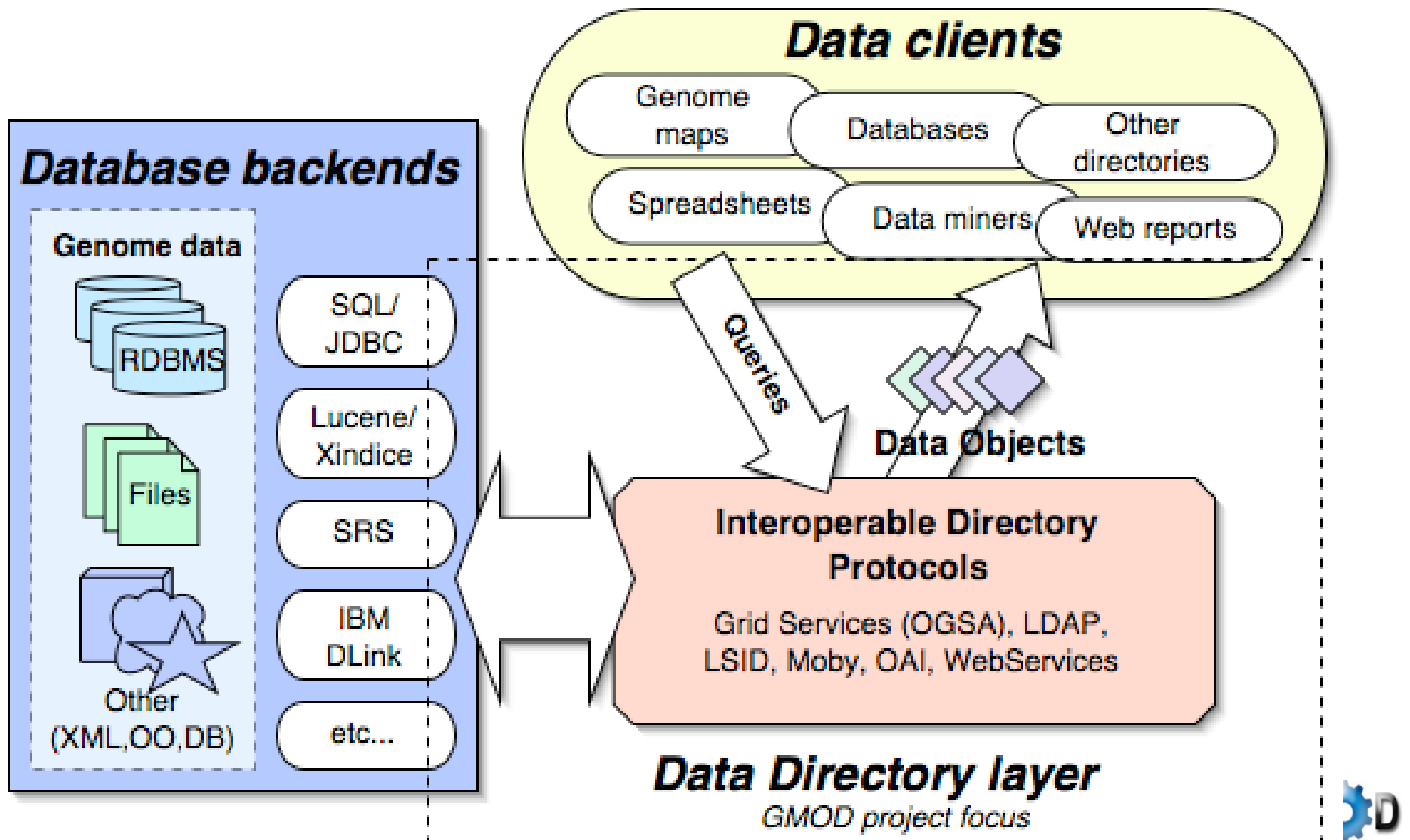
# Directories of Genome Data

---

- **Directories are a necessary step for bio grids**
  - *"broad and shallow"* directories federate the *"narrow and deep"* databases
- **Bio-Data Access Tools**
  - SRS, Sequence Retrieval System; Entrez ; AceDB; Genome relational databases (Ensembl, FlyBase, WormBase) ; IBM DiscoveryLink; BioDAS ; BioMoby
- **Directory services for data access**
  - Layer onto access tools for common query/retrieval
  - **LDAP**: mature, efficient for high volumes, query distributed directories ; works well with bio-access tools
  - **Web Services**: XML messages over Web ; wide industry support , standards are in progress



# Directory components





# Genome Directory Needs

---

- Build on **existing technology** for finding distributed objects
- **Efficient** for millions of objects, by the gigabyte and terabyte
- **Queries** distributed **across directories** of collaborating services
- **Support** existing and new **bioinformatics data access** (relational dbs, object and XML dbs, SRS, Entrez, AceDB)
- **Simple client** program **methods** for computable use of directories
- Flexible, **common schema** for describing objects
- **Replicate directories** and objects among bioinformatics centers
- **Peer-to-peer directories** for collaborative projects
- Strong **authentication and security** for data access





# Directory Standards

---

- Open Grid Services Architecture (OGSA)
  - SOAP based; query support for XML-SQL and Xquery.
  - Data Access project: <http://www.ogsa-dai.org.uk/>
- Lightweight Directory Access (LDAP)
  - Robust system for distributed search and retrieval
  - Object-centric, optimized for efficient read operations
  - Hierarchical, distributed and replicated in nature
- Life Sciences ID (LSID)
  - new standard for bio-object naming, with LDAP and WebServices implementations
- Moby project web services repository system





# Directory Tests

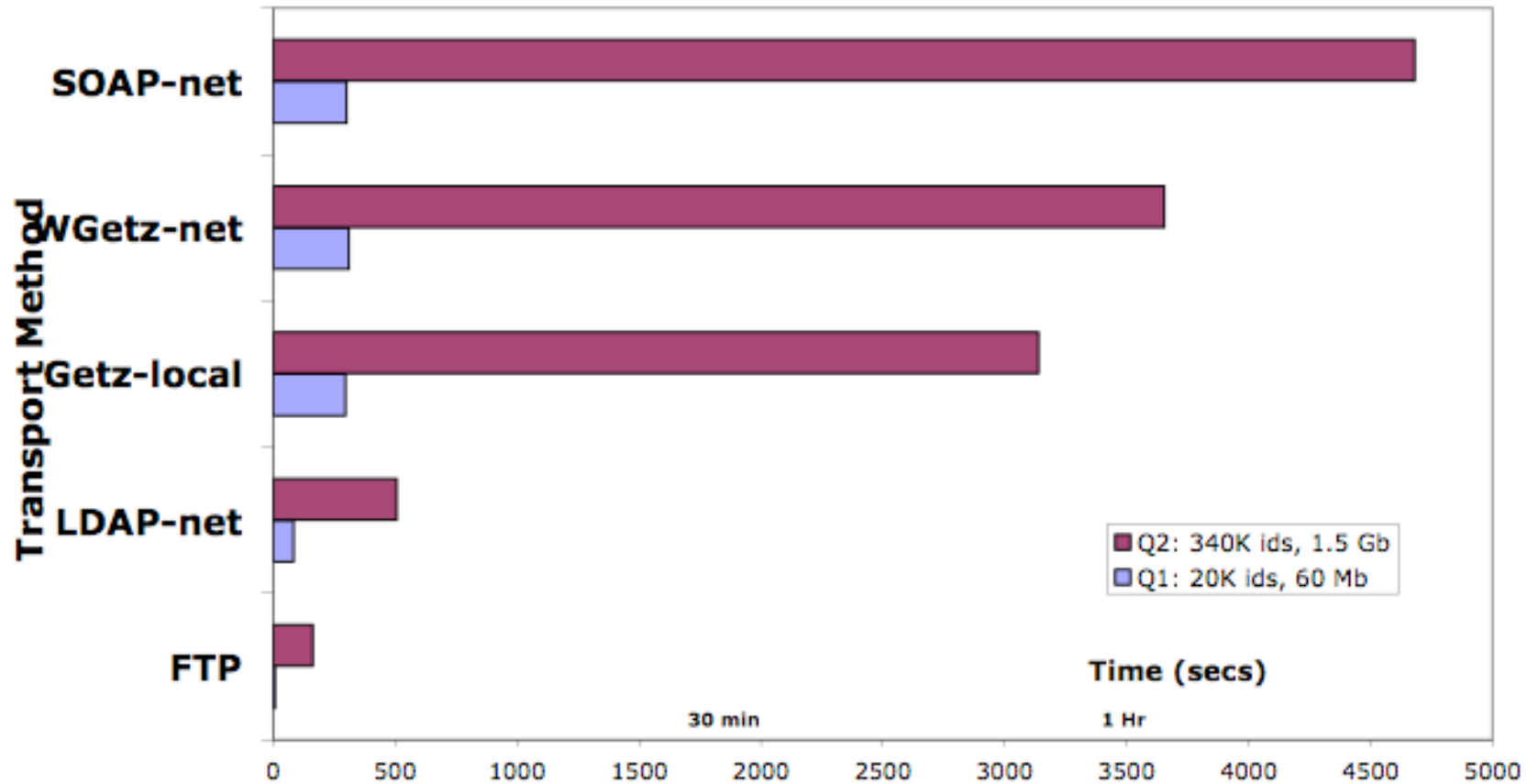
---

- Design and test distributed access with LDAP and Web Services
- SRS backend for efficient search/retrieval from GenBank, SwissProt/TrEMBL, LocusLink, Medline, many others
- Find & fetch 20,000 to 1.2 million objects
- LDAP is ~10x faster than WebServices
- Test in progress for FlyBase



# Directory Tests

**BioDirectory Search/Retreive Efficiency**



FTP > 5x LDAP > 5x Web

	FTP	LDAP-net	Getz-local	WGetz-net	SOAP-net
Q2: 340K ids, 1.5 Gb	162	506	3138	3653	4680
Q1: 20K ids, 60 Mb	8	84	297	309	300

Q1/Q2 - Query biosequence directories ; FTP - no query selection

Q1 = {swissprot trembl refseq}-des:kinase , 20K records; Q2 = genbank-org:drosophila , 340K records

gilbertd@bio.indiana.edu, Oct 2002





# Genome Directory Issues

---

*Directory tests at*

<http://iubio.bio.indiana.edu/biogrid/directories/>

