OGSA-DAI

NGS: Application Developer Training

23rd February 2007, eSI

open middleware
infrastructure institute uk
www.omii.ac.uk

Neil Chue Hong
EPCC
N.ChueHong@epcc.ed.ac.uk
+44 131 650 5957

Overview

- What is OGSA-DAI
- What you can do with OGSA-DAI
- How do you use OGSA-DAI on the NGS
- Workflow in OGSA-DAI
- What's coming up in OGSA-DAI v3.0?
- Where you can get more information

Data Service Challenges

Diversity — of data resource types, vendors, middleware, schema, metadata

Scale — of collections, formats, geographical, political and social distance

Ownership — on individual, group, and organisation levels; intersecting yet independent

Security — for client, service and data owner; at many levels, with many tradeoffs

OGSA-DAI

Use Cases for Data Services

- Data Filtering:
  - Single source producing large amounts of data distributed to many sites downstream
- Data Discovery:
  - many sources, many query entry points in a linked system
- Data Translation:
  - source to sink, conversion of data model / structure
- Data Federation:
  - many sources, linked to provide view as a single source
- Data Replication
  - full or partial copies to improve throughput
- Data Integration (model aggregation)
  - e.g. integration of time variant data, streams, files
- Data Integration (knowledge expansion)
  - forming links between databases to increase knowledge

OGSA-DAI

Data Service Spec Goals

Impose standard interfaces to:

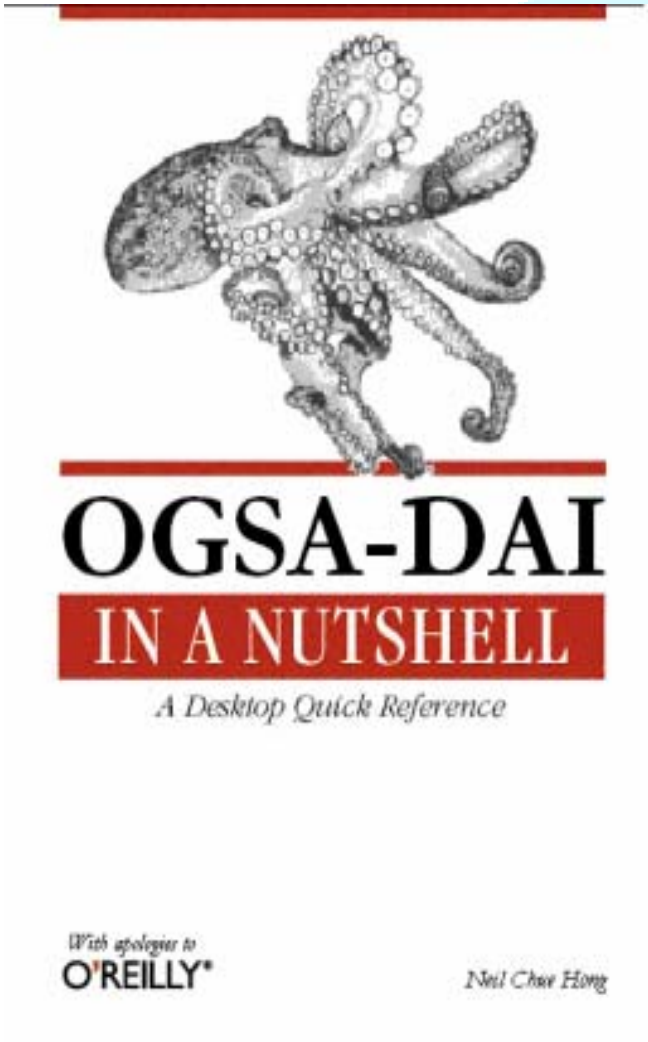
Make access transparent

Make integration easy

Make management simple

OGSA-DAI

OGSA-DAI In One Slide



Extensible

Portable

Easy to develop

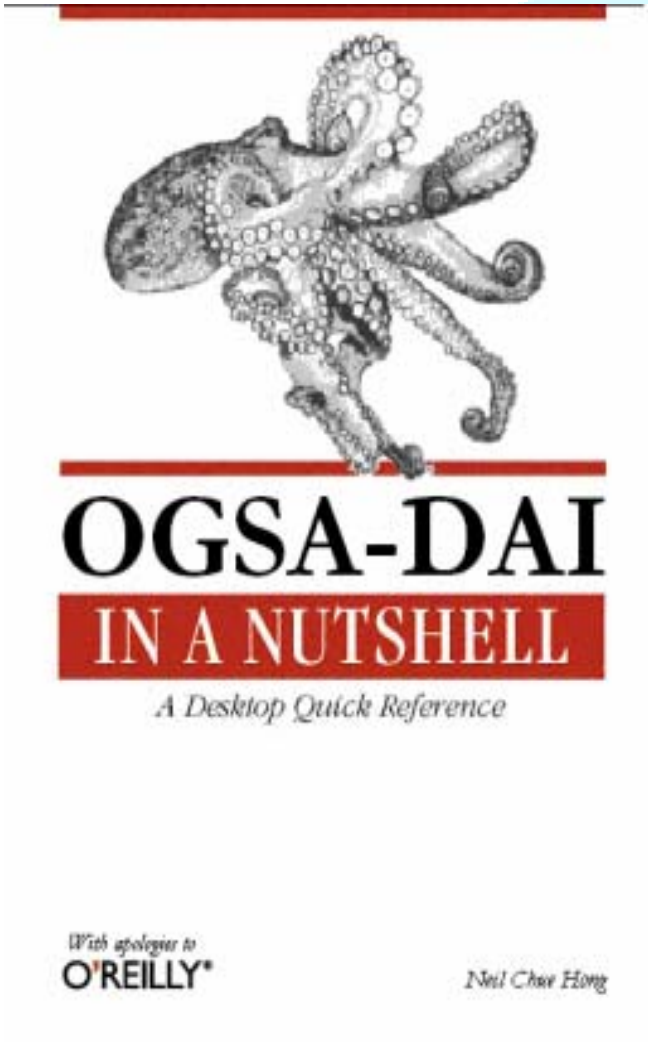
Diverse, independently curated data sources

We provide the generic

You develop the specific

OGSA-DAI

OGSA-DAI In One Slide



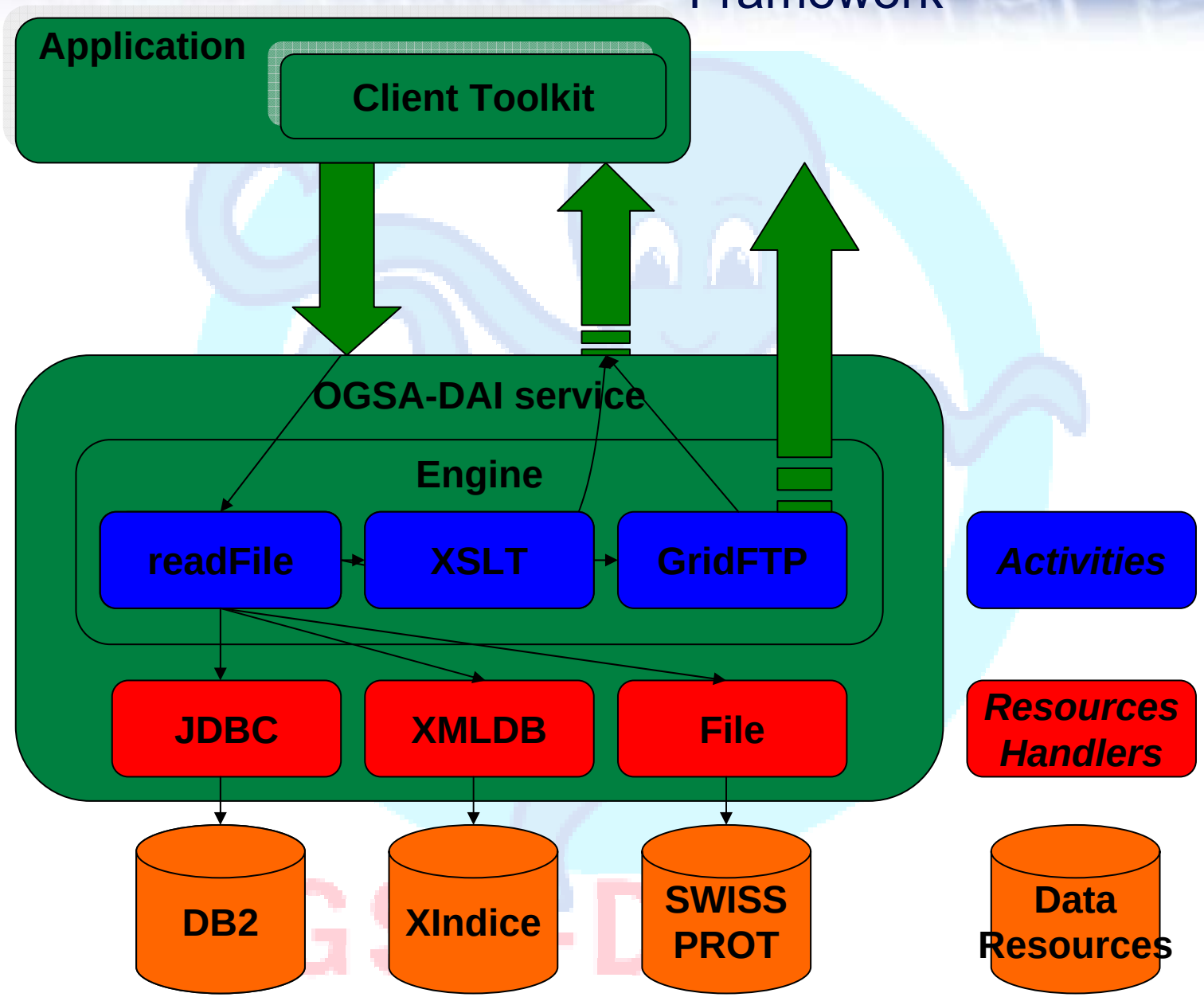
- **OGSA-DAI is a Java-based product that allows diverse, independently curated data resources to be exposed via Web services**
- **An *extensible framework* for data access and integration.**
- **Interact with data resources:**
  - **Queries and updates.**
  - **Data transformation / compression**
  - **Data delivery.**
- **Customise for your project using**
  - **Additional Activities**
  - **Client Toolkit APIs**
  - **Data Resource handlers**
- **Move computation to data**
- **A base for higher-level services**
  - **federation, mining, visualisation**

OGSA-DAI

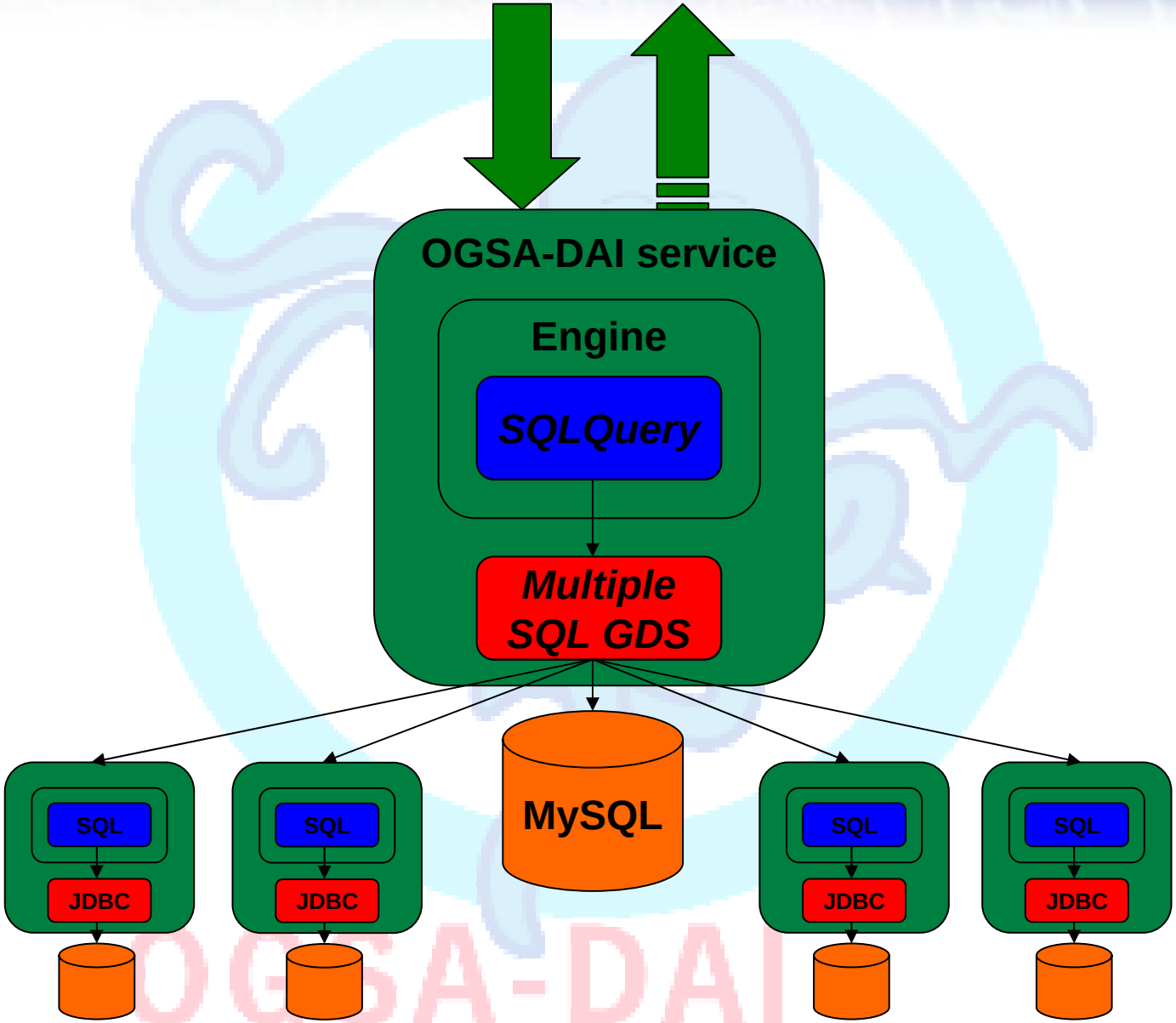
Overview

- What is OGSA-DAI
- What you can do with OGSA-DAI
- How do you use OGSA-DAI on the NGS
- Workflow in OGSA-DAI
- What's coming up in OGSA-DAI v3.0?
- Where you can get more information

The OGSA-DAI Framework



SQLBag Example

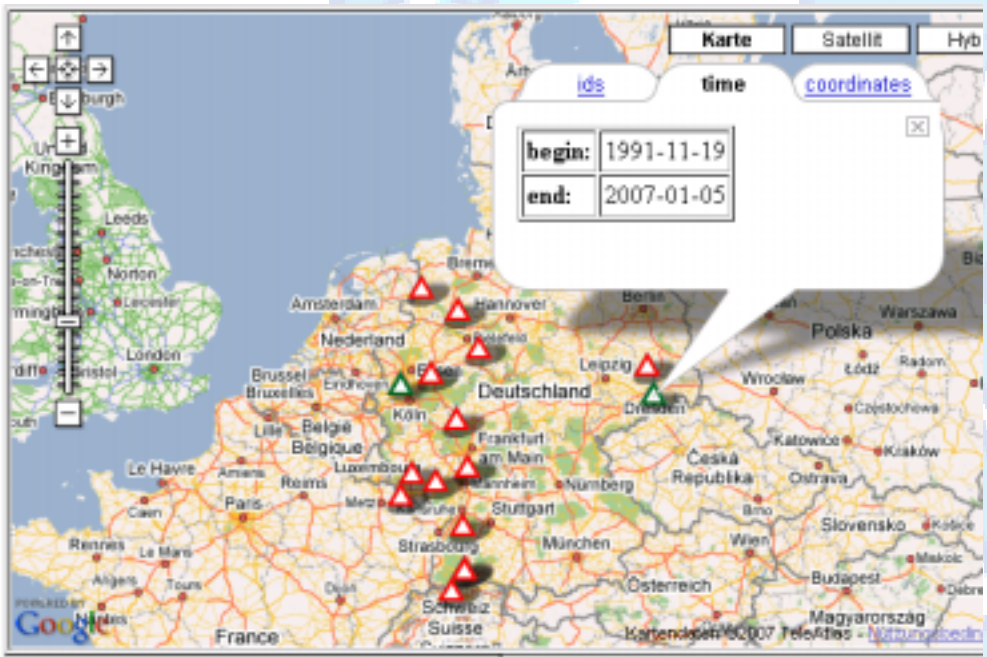


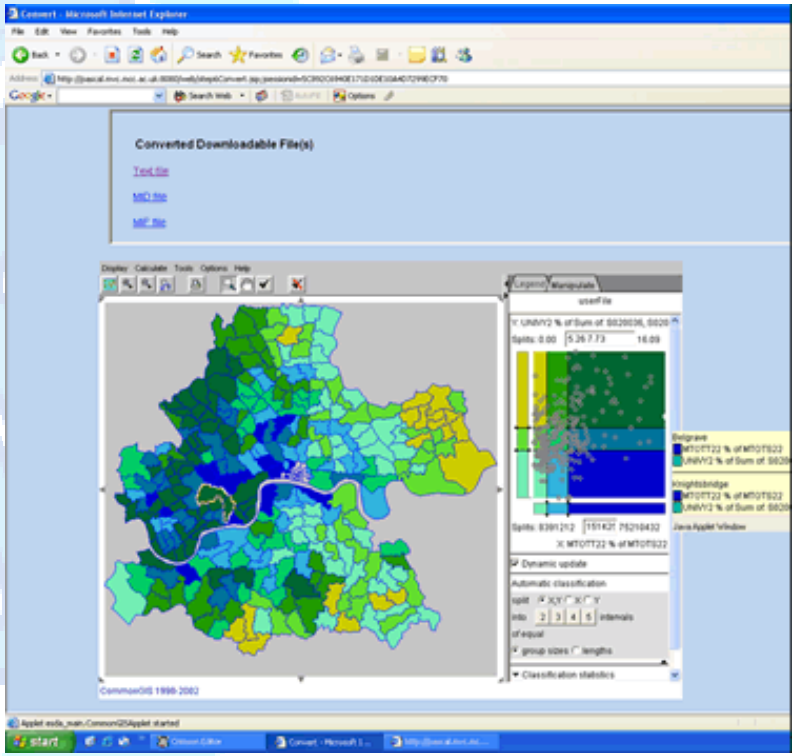
Core features of OGSA-DAI

- Data access, insert and update
  - Relational: MySQL, Oracle, DB2, SQL Server, Postgres
  - XML: eXist, XIndice
  - Files – CSV, BinX, EMBL, OMIM, SWISSPROT,…
- Data delivery
  - SOAP over HTTP
  - FTP; GridFTP
  - E-mail
  - Inter-service
- Metadata extraction
- Data transformation
  - –XSLT
  - –ZIP; GZIP
  - –Projections
- Security
  - –X.509 certificate based security
- Multi OS support
  - –Java 1.4/1.5 based
- Client API
- Documentation/ Tutorials

OGSA-DAI

Bringing together PUBLIC and PRIVATE data

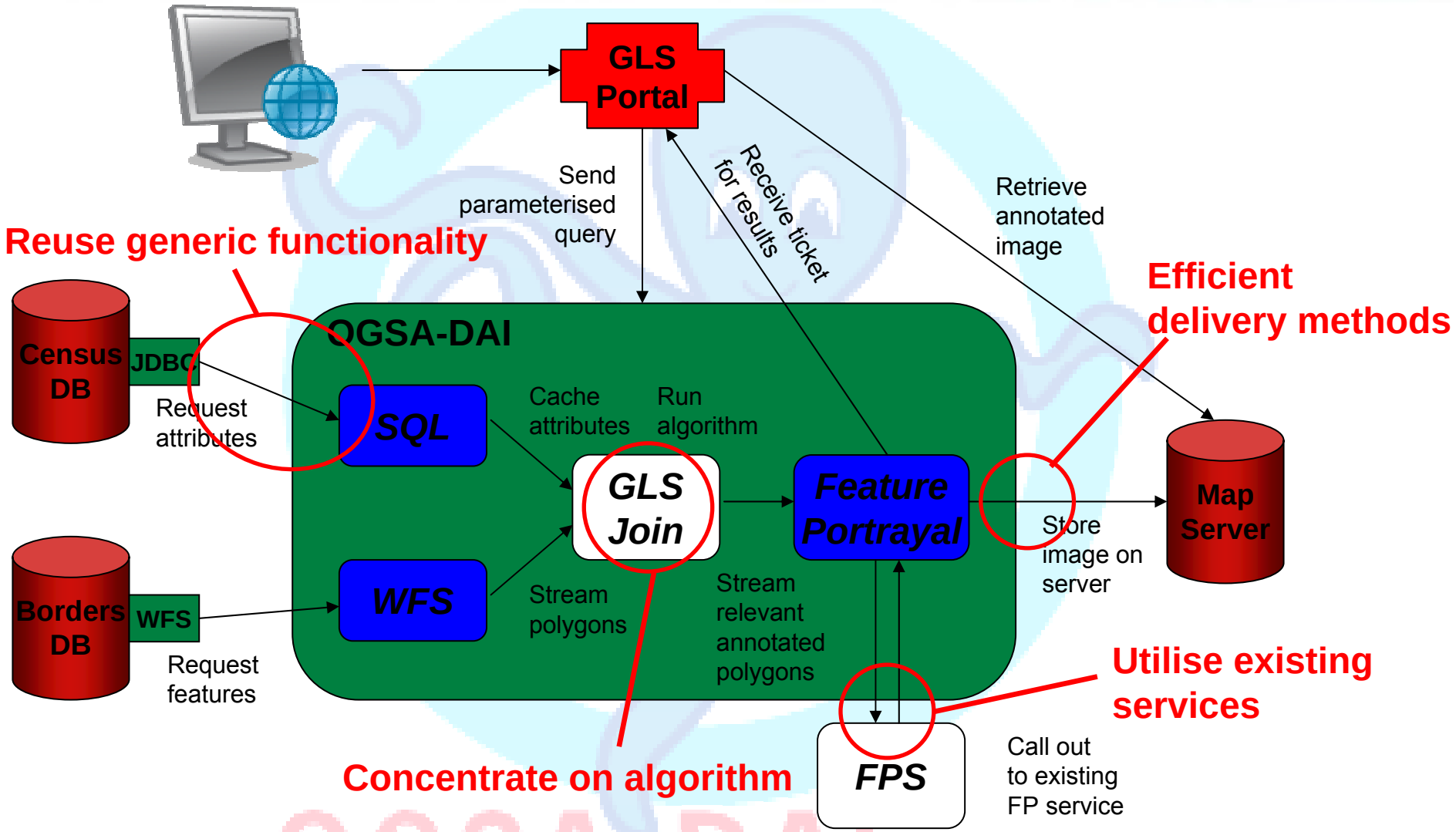




OGSA-DAI

Images from UNIDART and ConvertGRID projects

Demographic forecasting



OGSA-DAI

Other common patterns

- Robust vertical integration (eDIAMOND, GeneAssociationAnalysis)
  - the more results you can get, the better confidence interval you have
- Content-based retrieval (AHDS)
  - queries are based on performing some computation/querying within the initial result set, e.g. for images, videos
- Shared annotation systems
  - common data sets, personal annotations
  - annotations shared and queryable for greater relevance

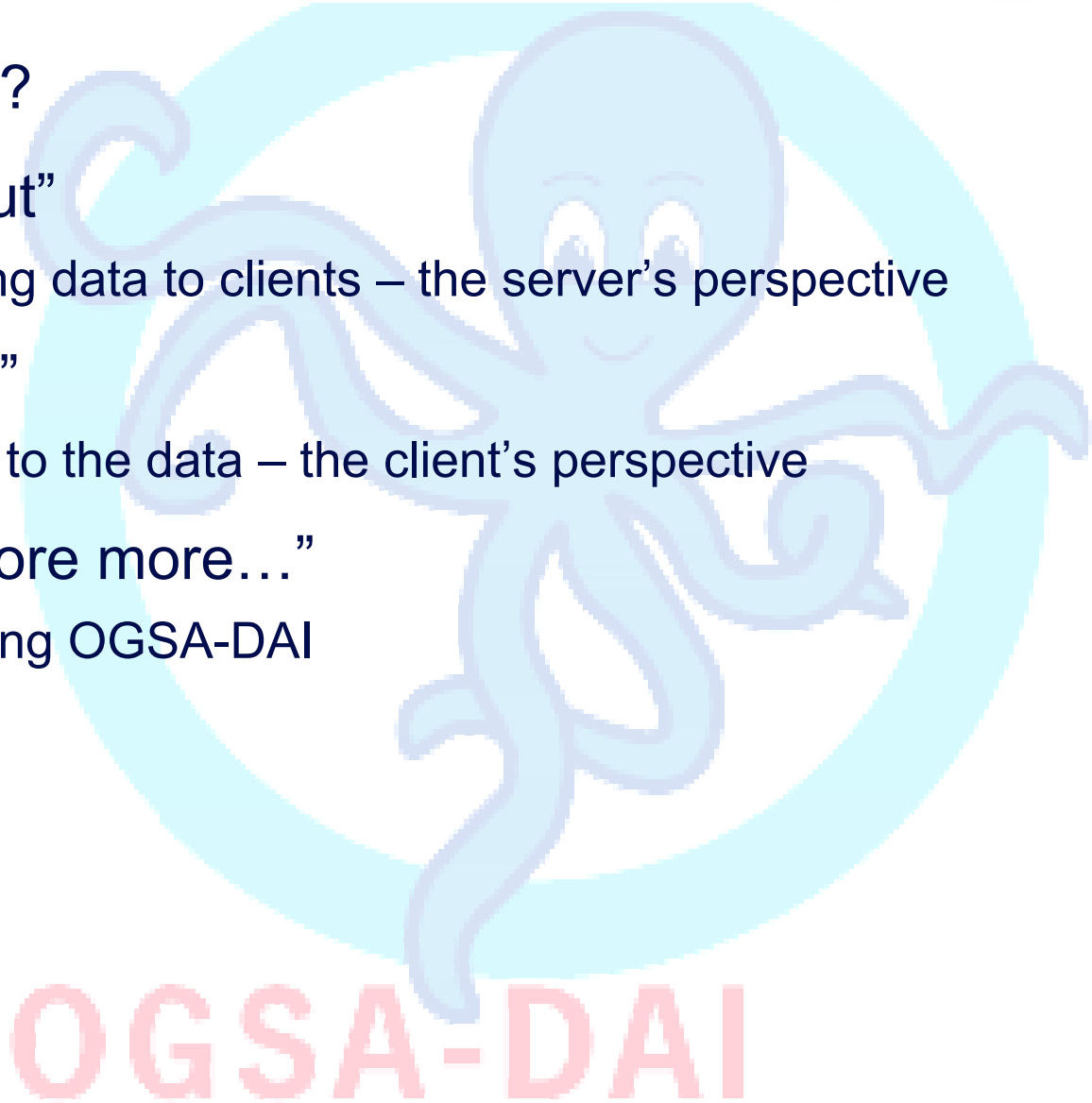
OGSA-DAI

Overview

- What is OGSA-DAI
- What you can do with OGSA-DAI
- How do you use OGSA-DAI on the NGS
- Workflow in OGSA-DAI
- What's coming up in OGSA-DAI v3.0?
- Where you can get more information

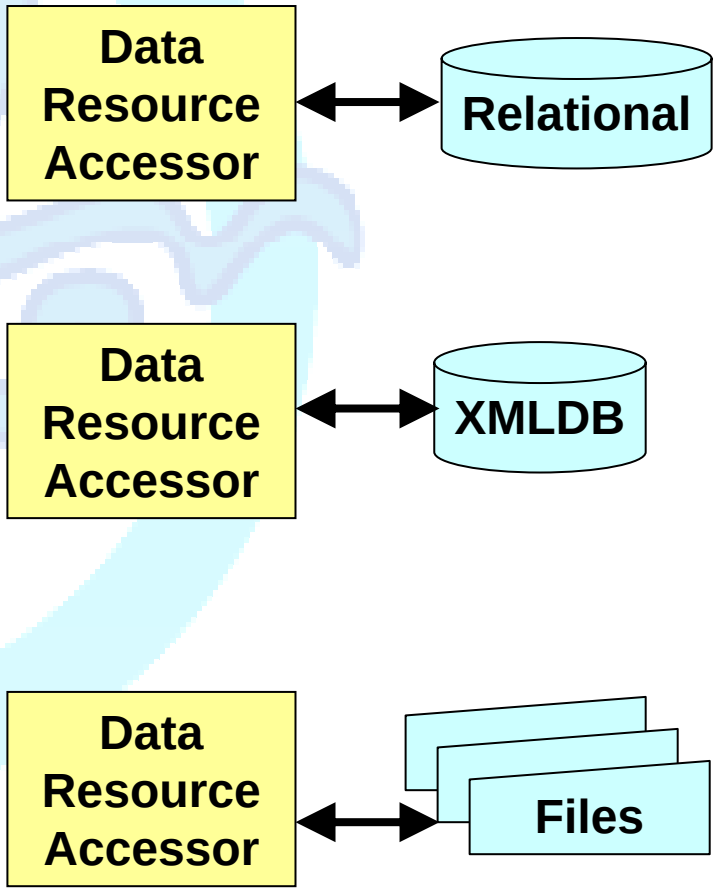
Outline

- What is it?
- “Let us out”
  - Exposing data to clients – the server’s perspective
- “Let us in”
  - Getting to the data – the client’s perspective
- “More, more more…”
  - Extending OGSA-DAI

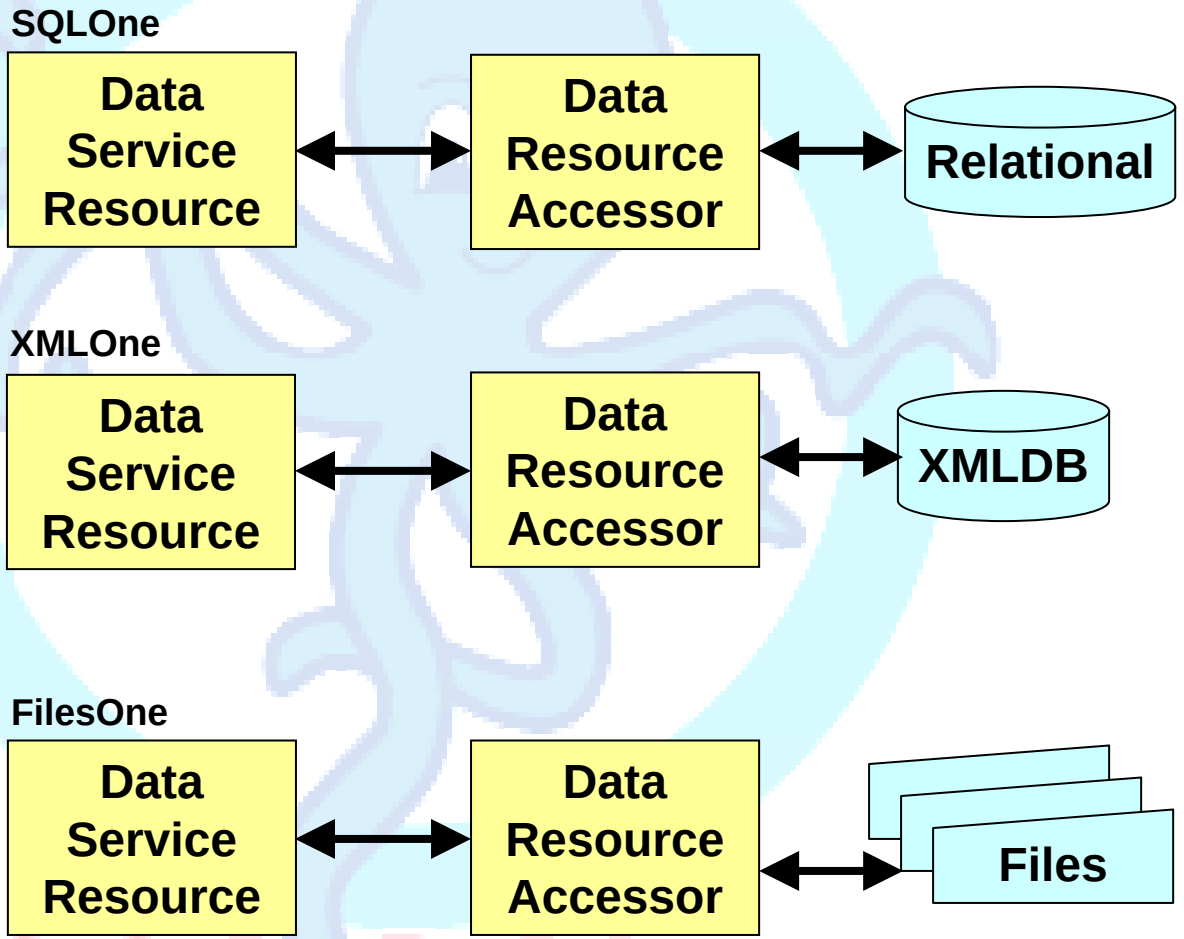


Data resource accessors

- Interfaces between data resources and OGSA-DAI
- Relational
  - JDBC drivers
  - `java.sql.*`
- XML
  - XMLDB API and compliant drivers
  - `org.xmldb.api.*`
- File system
  - Java file and directory utilities
  - `java.io.*`



Data service resources



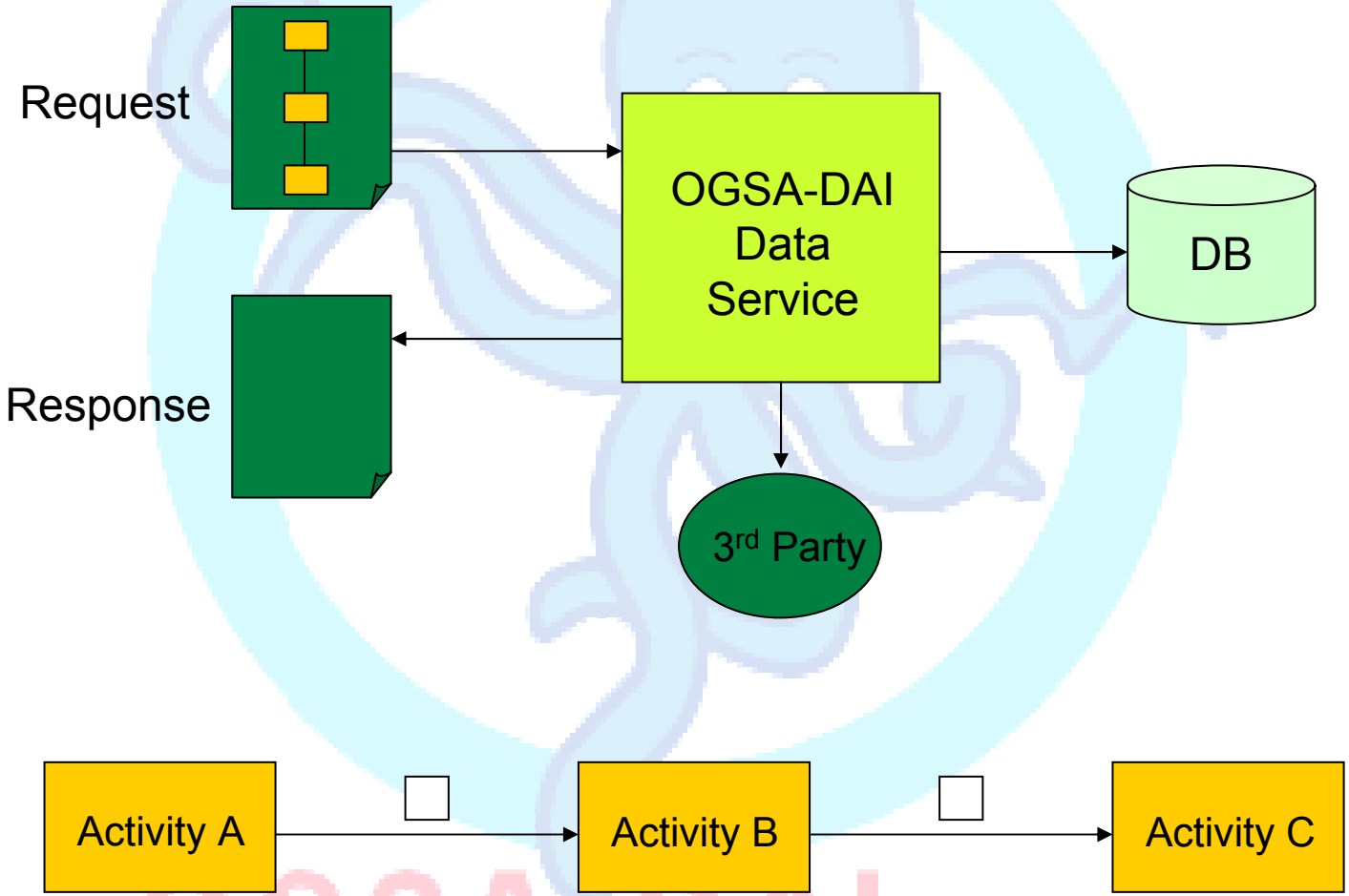
OGSA-DAI

Data service resources

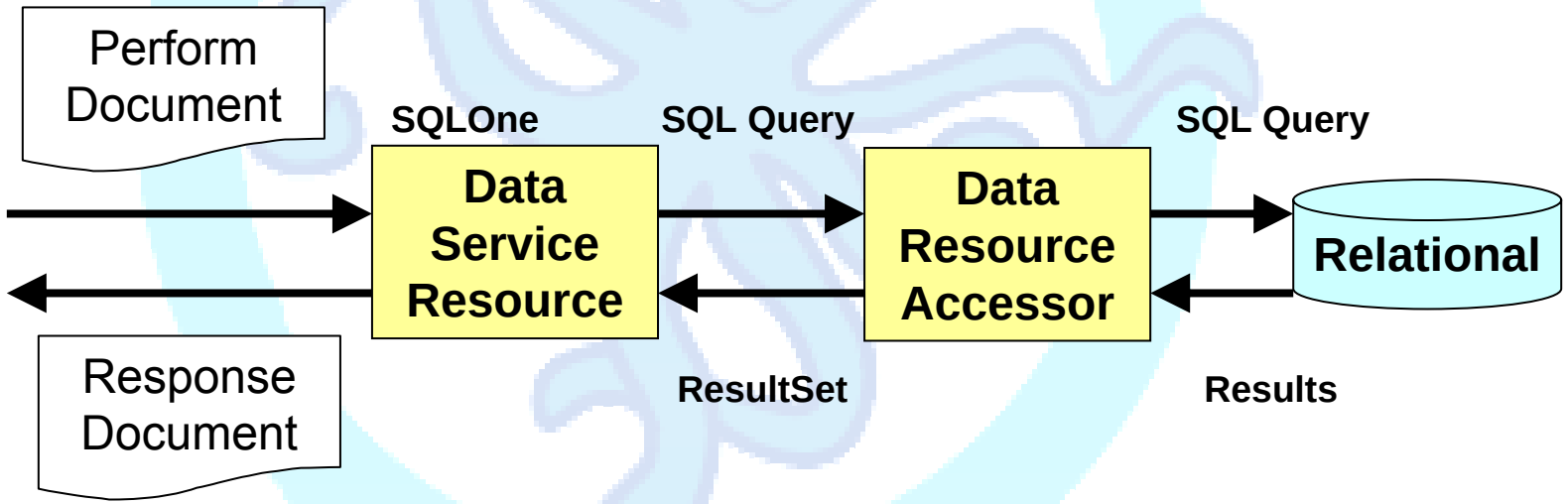
- OGSA-DAI's core functionality
- Manages
  - Access to a data resource via a data resource accessor
  - Execution of data-related activities
  - Data caching and streaming of data to and from clients
  - Creation, access and termination of sessions
- Exposes data service resource properties
  - Information about a data resource
  - Information about supported activities
  - Information about current requests

OGSA-DAI

OGSA-DAI Request/Response



Requests and responses

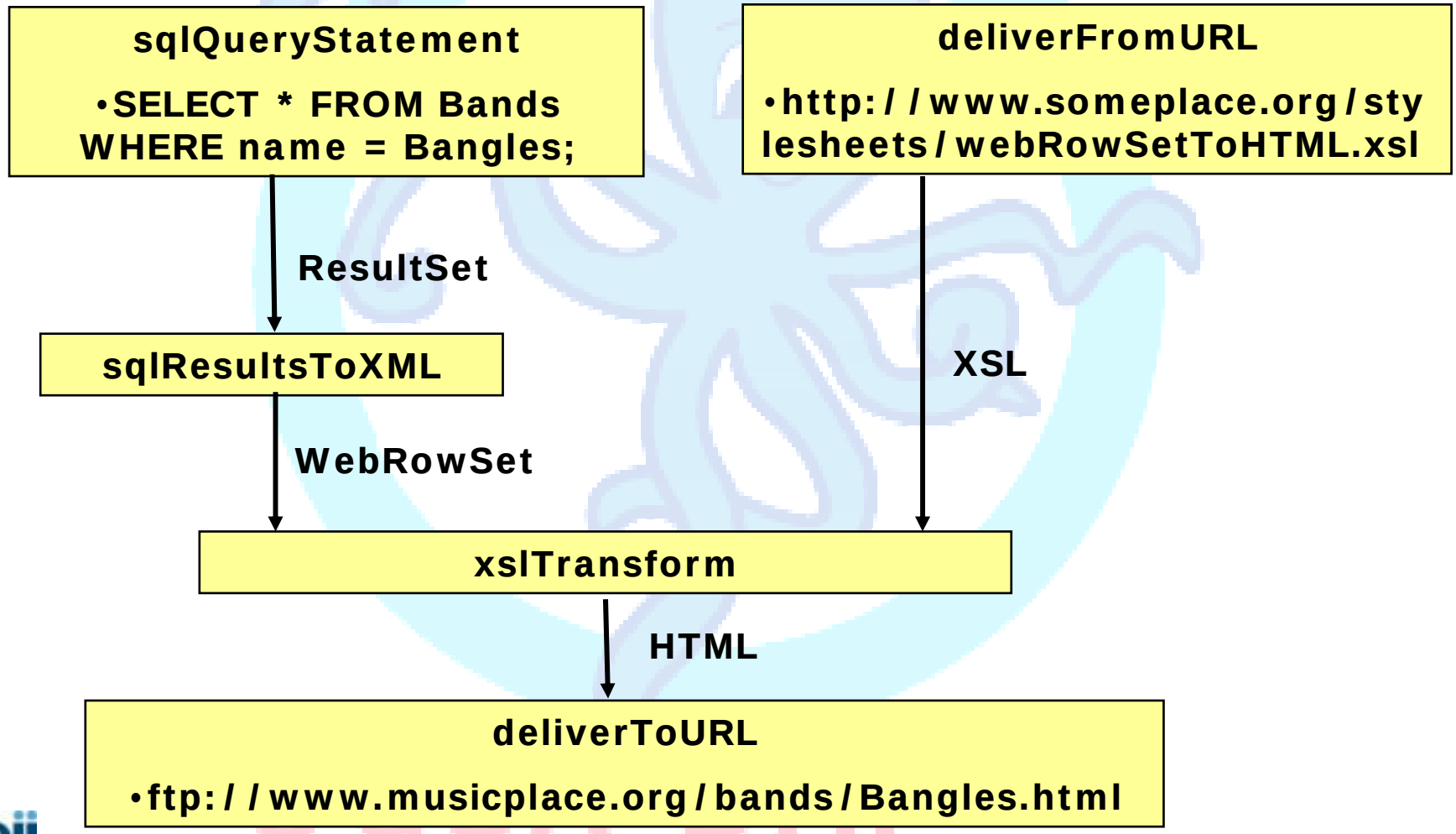


OGSA-DAI

Requests and responses

- Request
  - A connected collection of activities that the data resource executes
  - Flow control – sequential or parallel execution of activities
  - XML perform document submitted by a client
- Activity
  - An individual data-related operation
  - 0 or more inputs and 0 or more outputs
- Response
  - Status of execution of a request possibly with result data
  - XML response document returned to a client
- OGSA-DAI engine
  - Parses requests, executes activities, builds responses

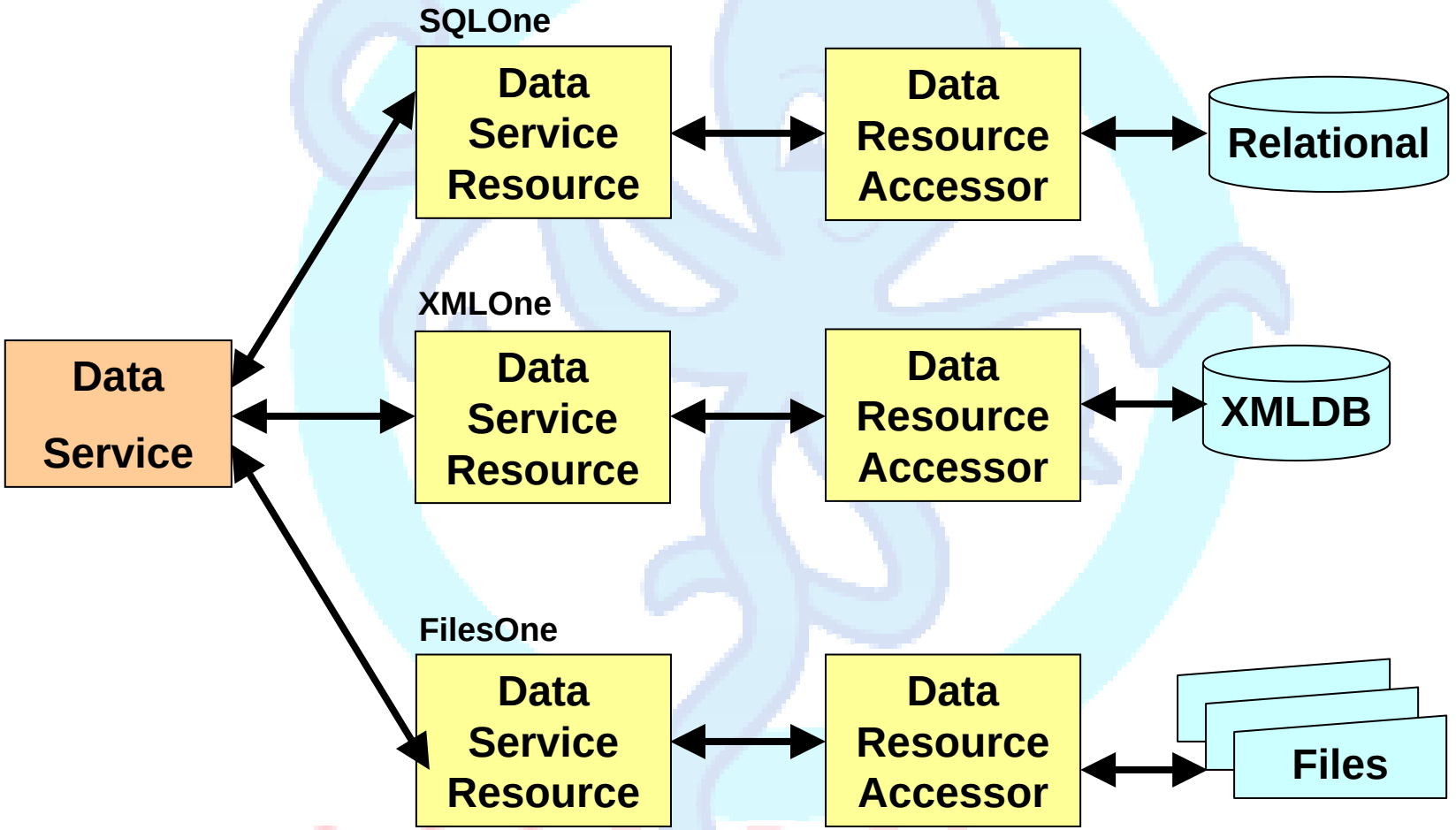
Example OGSA-DAI Request



Activities

- Resource-specific
  - Relational
  - XMLDB
  - Files
  - Multi-resources
- Transformation and compression
- Delivery
- Resource creation and destruction

Data services

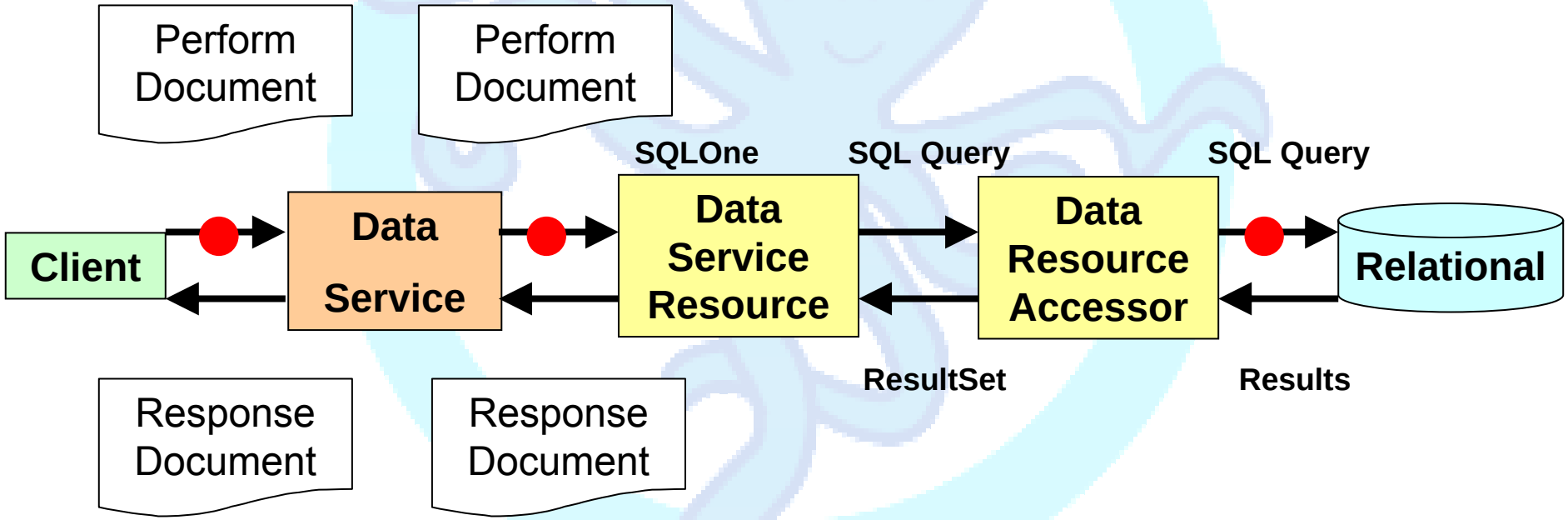


OGSA-DAI

Data Services

- Web services
  - Expose 0..N data service resources to the outside world
- Two flavours
  - OGSA-DAI WSRF services
    - Compliant with the Web Services Resource Framework
    - Implemented using Globus Toolkit (4.0+)
  - OGSA-DAI WSI services
    - Compliant with vanilla WSDL
    - Implemented using Apache Axis (1.2.1 or 1.2RC3)

Clients

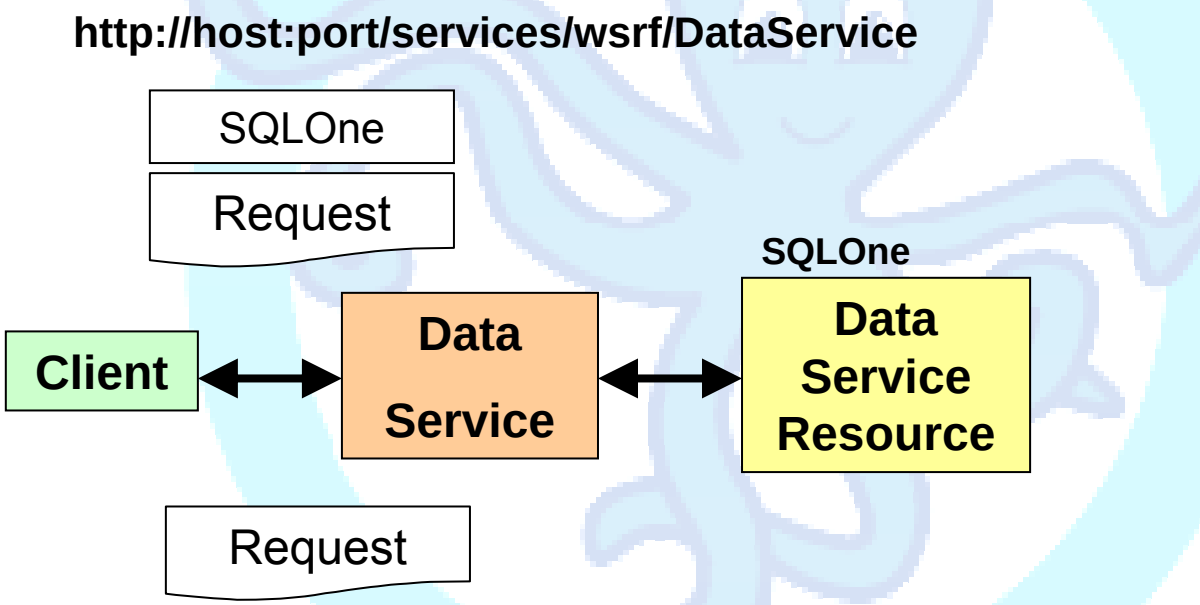


Also able to perform per activity authz

Authorization points

OGSA-DAI

Identifying a data service resource

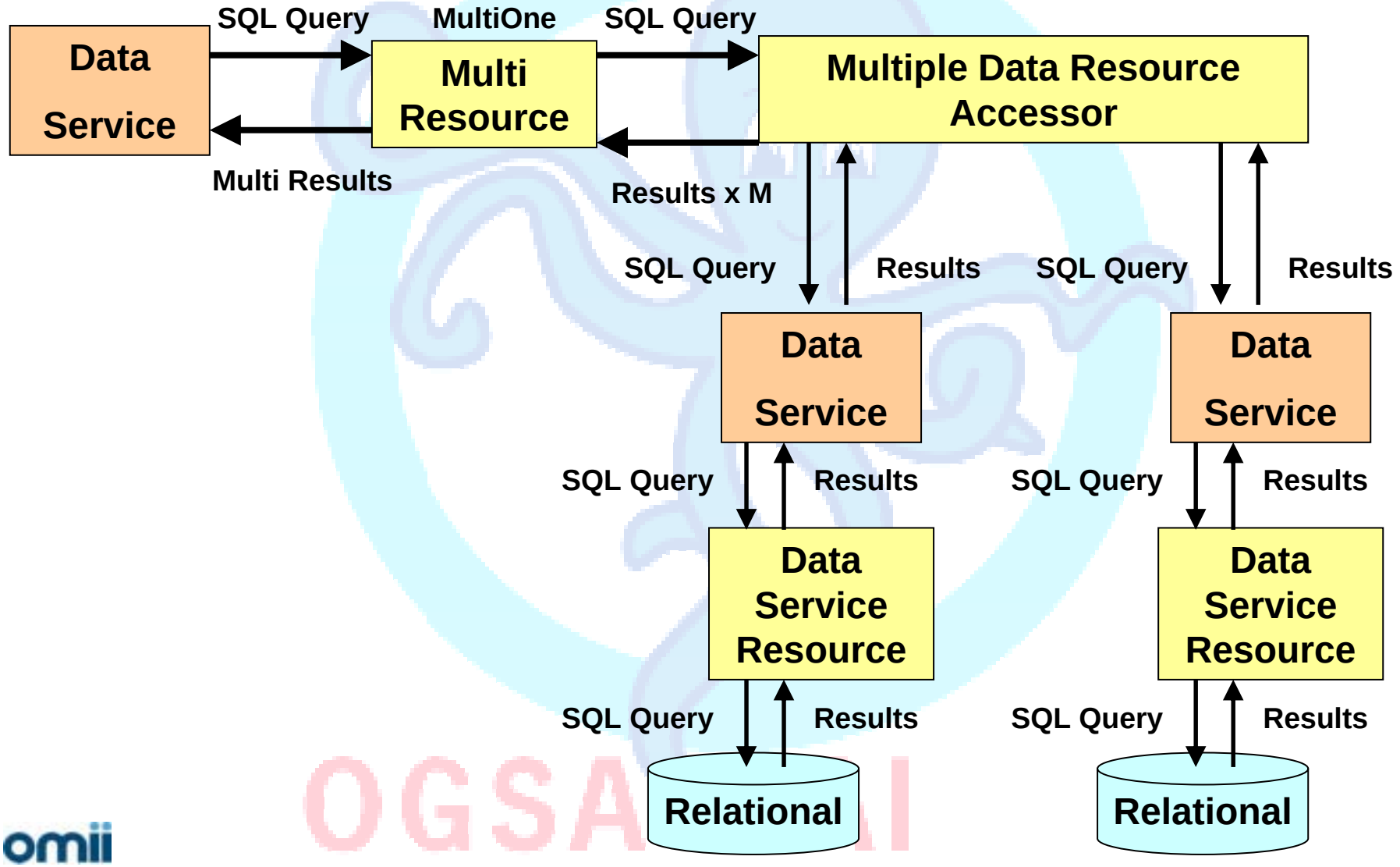


OGSA-DAI

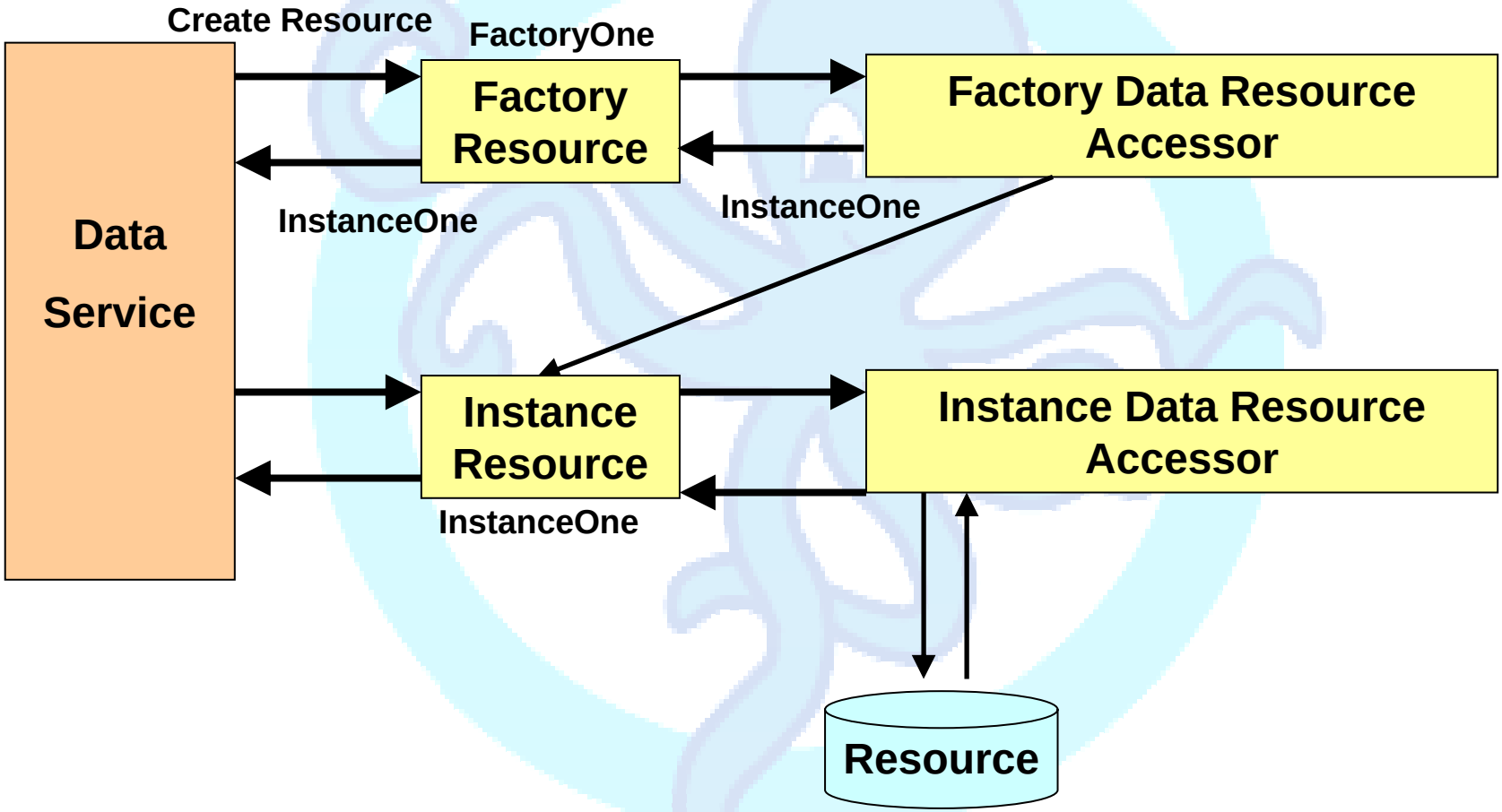
Clients and the client toolkit

- Clients interact with data services via SOAP over HTTP
  - Deduce service interface from service WSDL description
  - Construct SOAP request to invoke operation
  - Parse SOAP response from service
  - Resource identification scheme must be assumed from WSDL namespace
- OGSA-DAI client toolkit:
  - Construct and submit requests in Java not XML
    - Toolkit handles SOAP request construction and response parsing
  - Renders OGSA-DAI service types transparent
  - Java abstractions of
    - Data services
    - Data service resource IDs and session IDs
    - Requests and responses
    - Activities

Relational multi-resources



Factory resources



OGSA-DAI

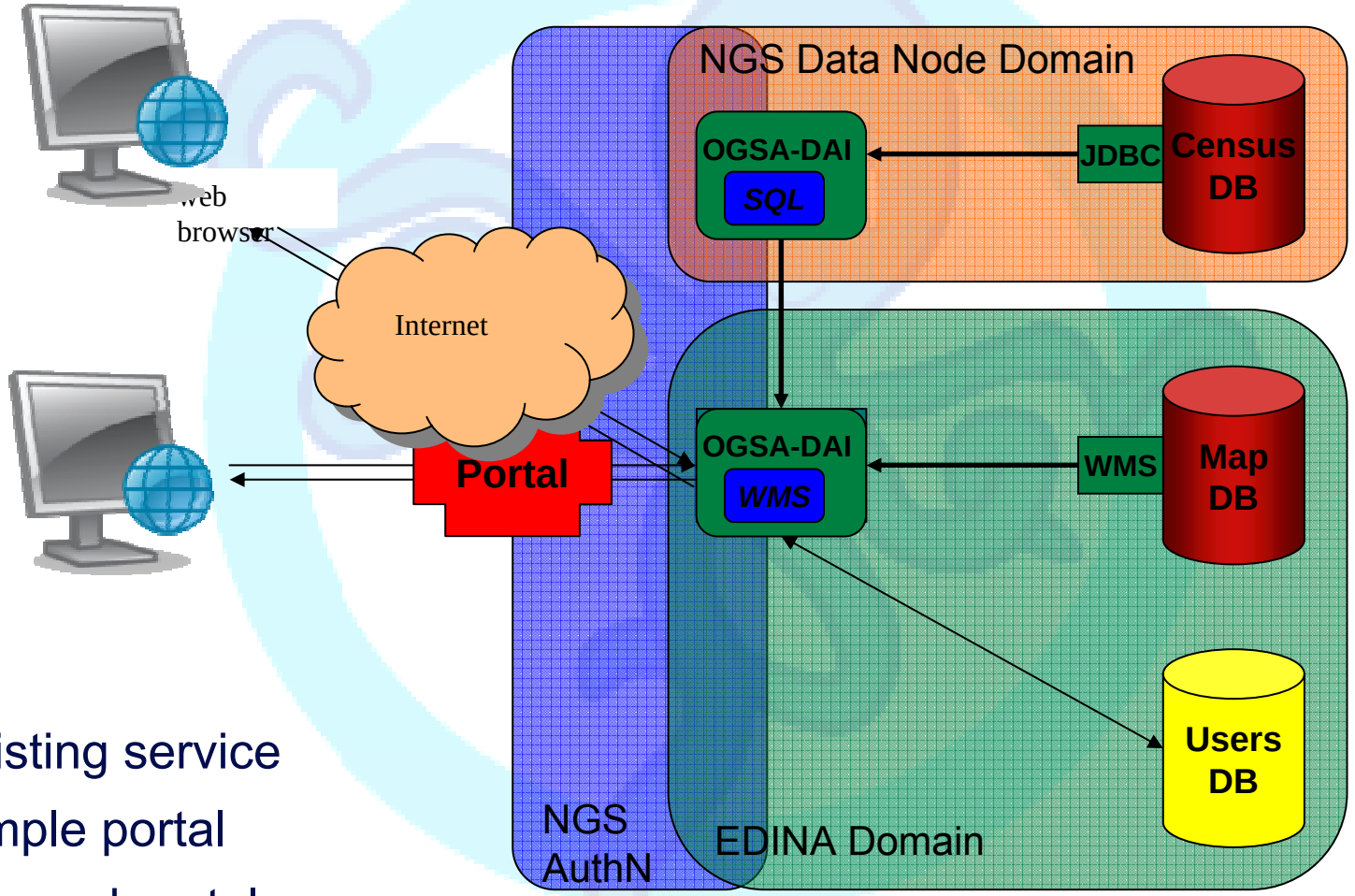
Extending OGSA-DAI

- Application-specific data resource accessors
  - Expose local or remote data resources
  - Expose virtual resources created by aggregation or integration
  - Create/destroy of persistent/transient data service resources
- Application-specific activities
  - Can be resource specific e.g query or update
  - Or generic e.g. transformation, compression, delivery, resource management, monitoring
- Application-specific authorization
  - Resource access
  - Activity execution

OGSA-DAI and the NGS

- 1) Host data in NGS Oracle services and use OGSA-DAI to expose them via the data nodes
- 2) Use OGSA-DAI clients on compute nodes to gather data from remote data sources for applications running on the compute nodes
- 3) Use OGSA-DAI services on the compute/data nodes to store data generated by application on the compute nodes
  - Security and provenance
  - Staging and transfer
  - It's important to make sure you are doing the sensible thing with your data! *compute to data or data to compute?*

Example: Map Retrieval

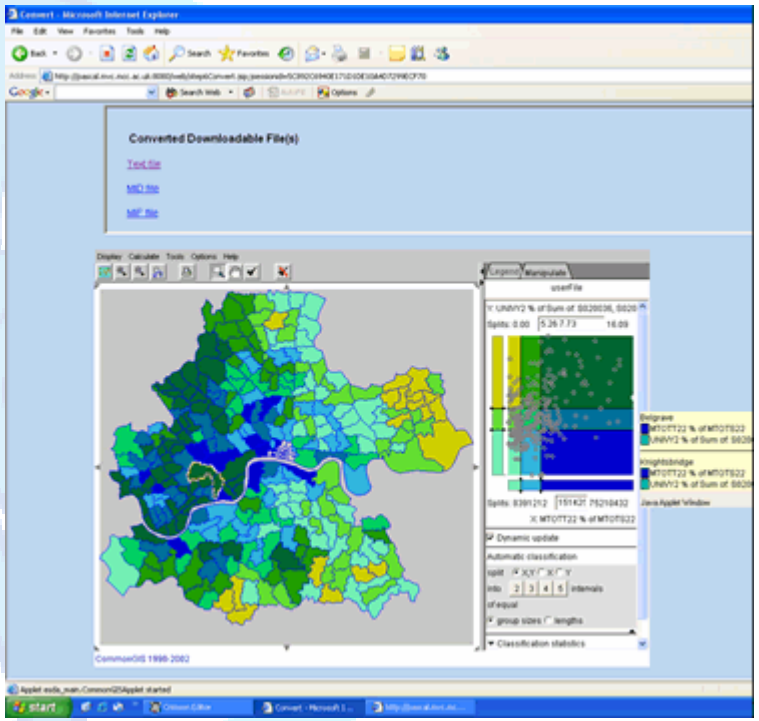


1) Existing service
2) Simple portal
3) Secured portal
4) Integrated portal

OGSA-DAI

ConvertGrid

- Facilitate comparison of data for different geographies
  - exploit ONS All Fields Postcode Directory
  - maps well to many other census, health, electoral etc. geographies
  - convert from source geography to a chosen target geography
  - achieves single signon using Athens
  - provide simple API using OGSA-DAI
- http://www.sve.man.ac.uk/Research/AtoZ/ConvertGrid/



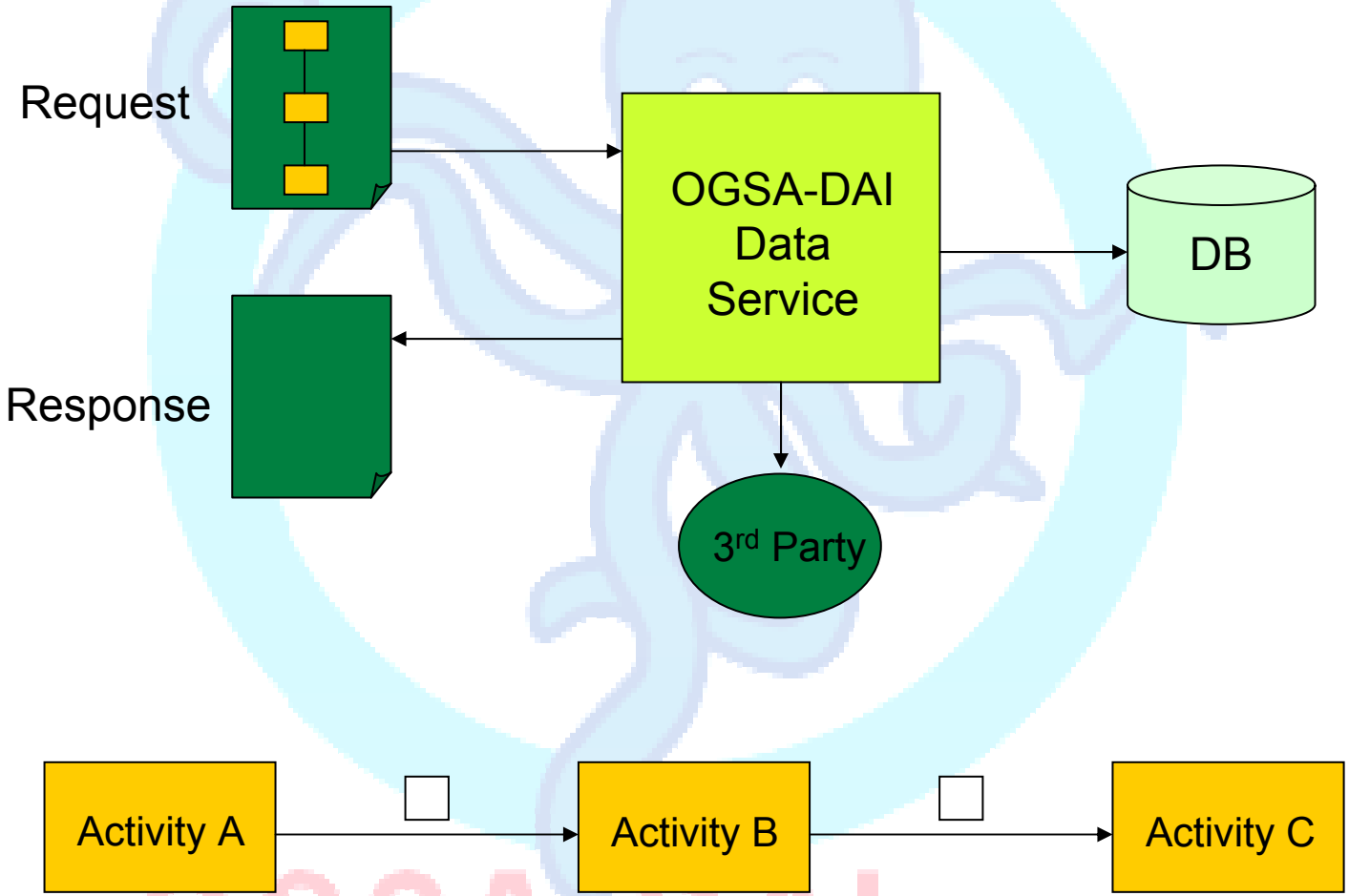
Relationship between house prices (Experian dataset) and the percentage of young people in the 16 -19 age group entering University (ONS datasets) for the London area (at 1991 Census ward level)

OGSA-DAI

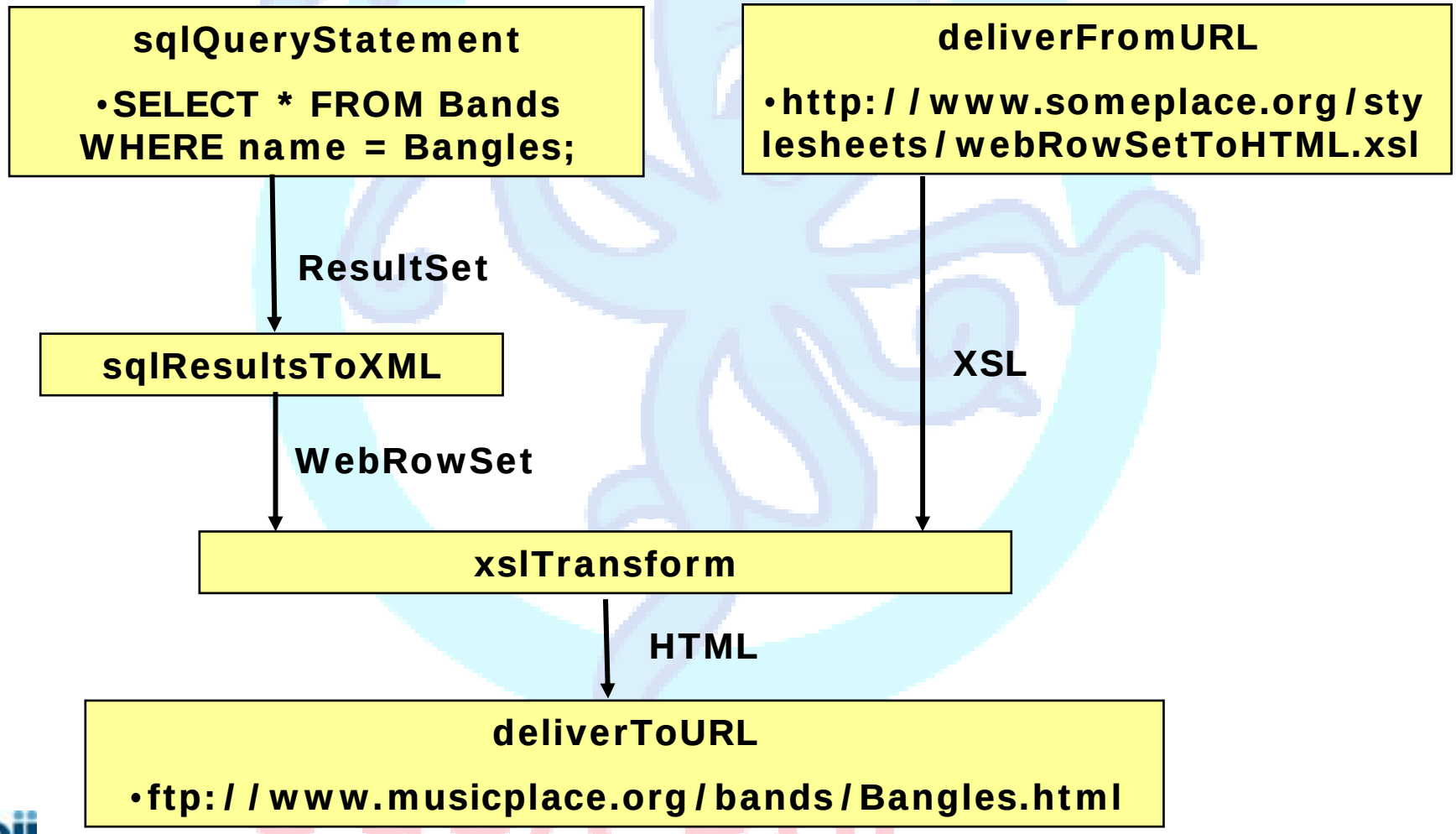
Overview

- What is OGSA-DAI
- What you can do with OGSA-DAI
- How do you use OGSA-DAI on the NGS
- Workflow in OGSA-DAI
- What's coming up in OGSA-DAI v3.0?
- Where you can get more information

OGSA-DAI Request/Response

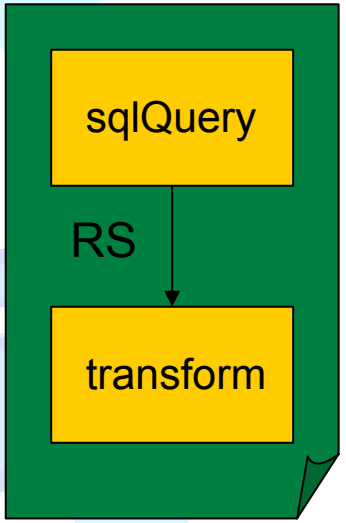


Example OGSA-DAI Request



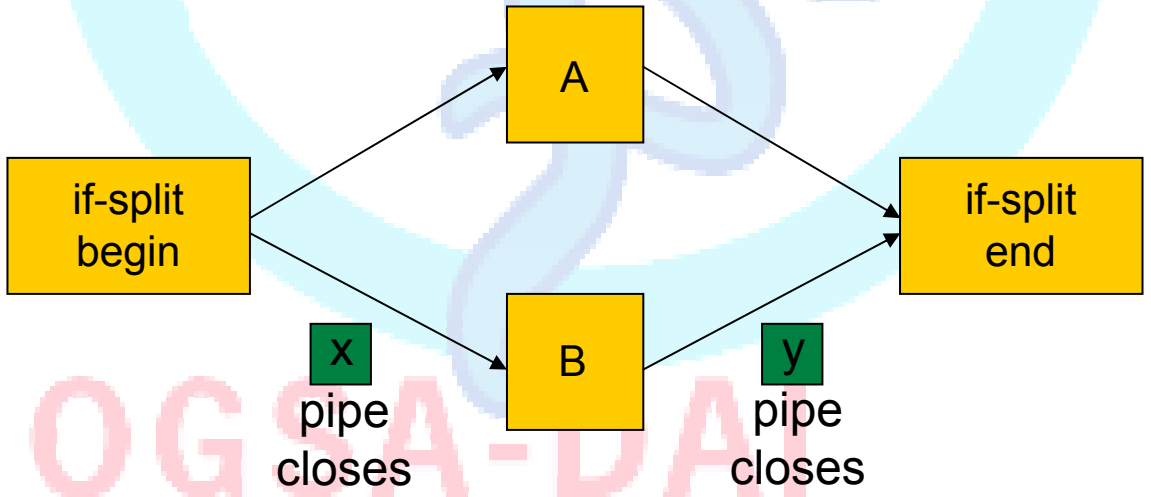
Workflow Distinctive Features (1)

- Keep everything local
  - –We have homogeneity
  - –Avoid data movement
    - –Call with whole result
    - –Separate calls
  - –E.g. we can pass object references
- Activities are configured at service deployment
  - –Cannot use any “service” you want in your workflow



Workflow Distinctive Features (2)

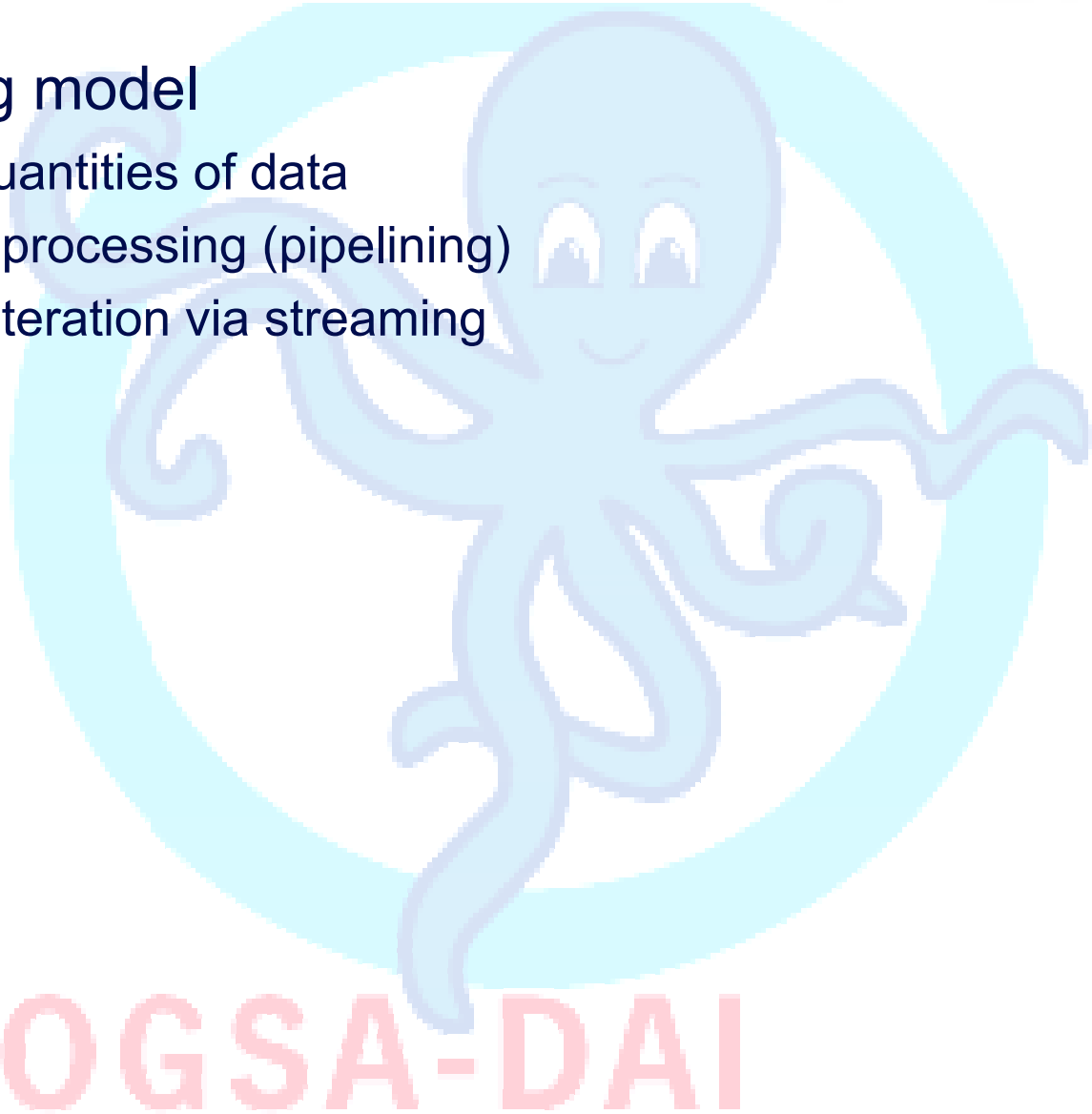
- Simple, efficient workflow language
  - –Sequence, flow
  - –Fits our data processing needs
  - –Other constructs can be used at the activity level (e.g. exclusive choice – a.k.a. if-split)



OGSA-DAI

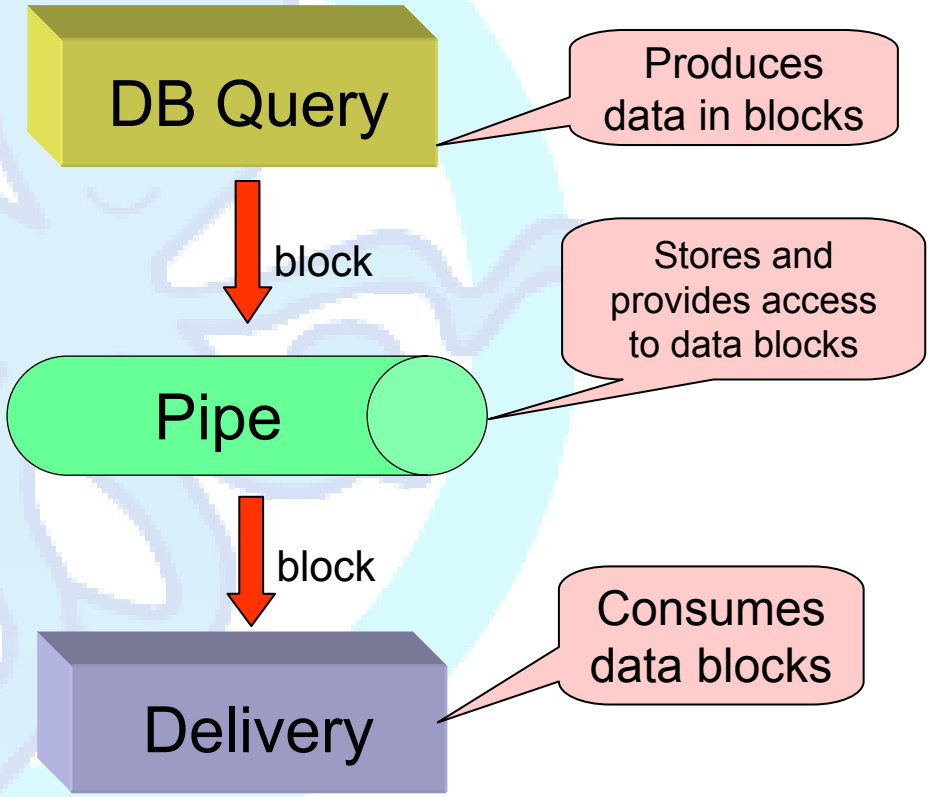
Workflow Distinctive Features (3)

- Streaming model
  - Large quantities of data
  - Parallel processing (pipelining)
  - Implicit iteration via streaming



Enactment Model

- A.k.a. Activity Framework
- Core component of OGSA-DAI
- It is responsible for performing tasks (activities) and streaming data



OGSA-DAI

Activity Processing

OGSA-DAI 2.2

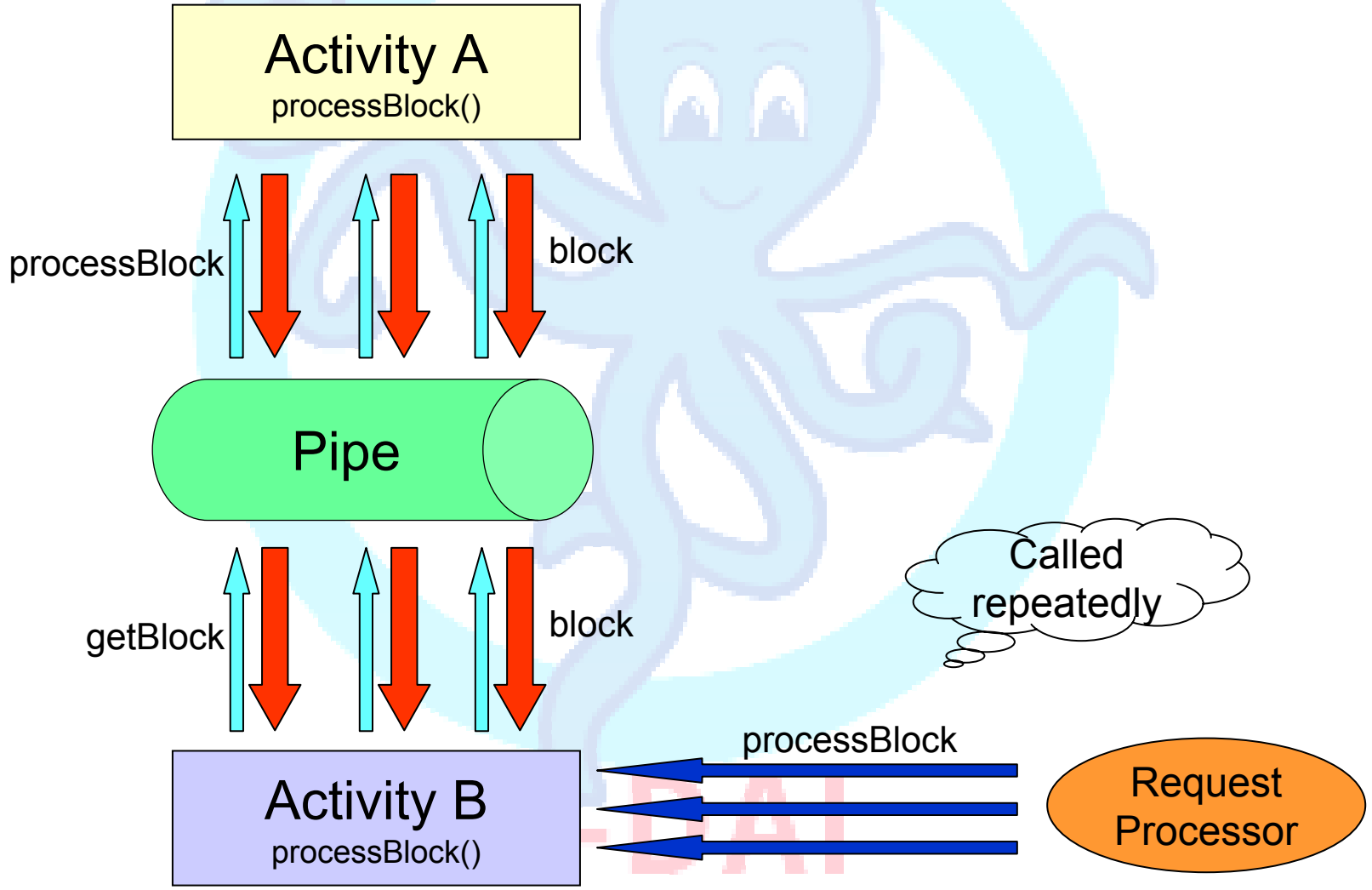
- Processing of blocks (and therefore activities) is controlled by the pipe – from outside the activity
  - –**processBlock()** is called many times until processing is complete
  - –Usually consumes and produces a single block per call

OGSA-DAI 3.0

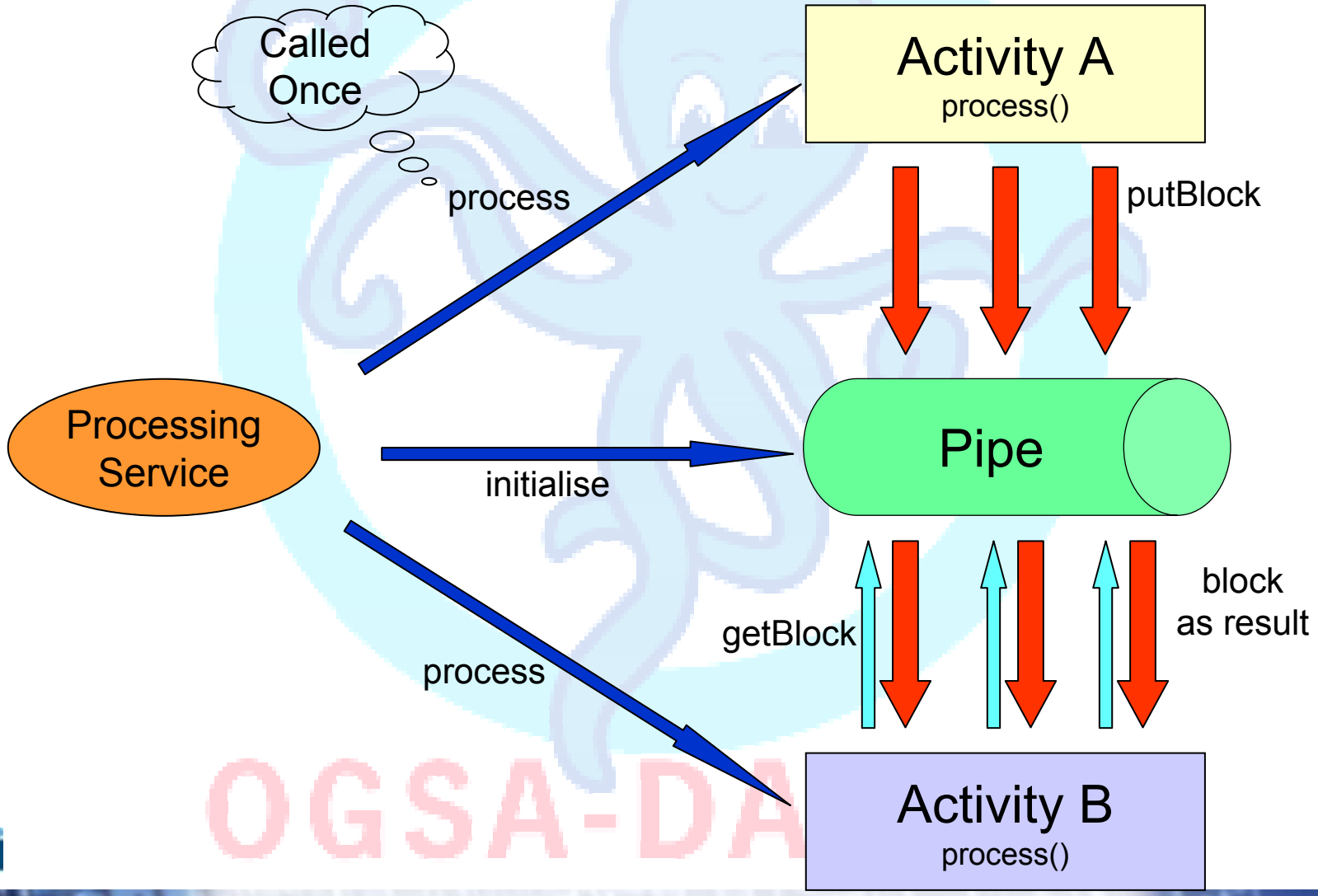
- Processing of blocks will be controlled by the activity
  - –**process()** is called exactly once
  - –Consumes and produces blocks as necessary
  - –Each activity in a pipeline processes within each own thread
  - –Pipes receive and may buffer blocks until they are requested

OGSA-DAI

Current Model



New Model



OGSA-DA

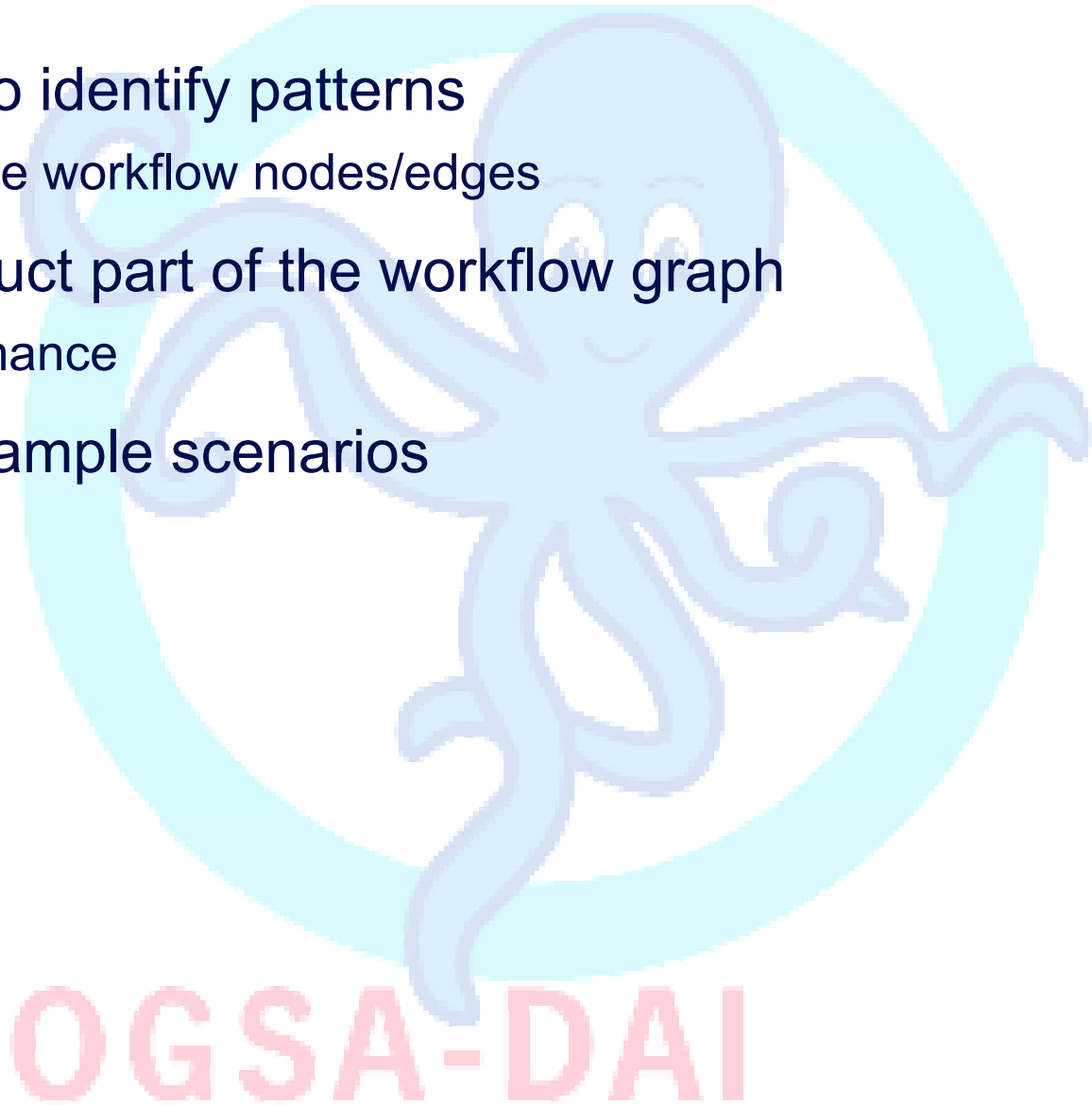
Improvements to the Enactment Model

- Each activity is run in its own thread
  - Rather than one thread per chain
- Each activity is called to process only once
  - Rather than once per block
- Each activity sends its result to the pipe when ready
  - Rather than wait to be called
- The pipe now buffers
  - Rather than delegate the request for new block
- Performance improvements

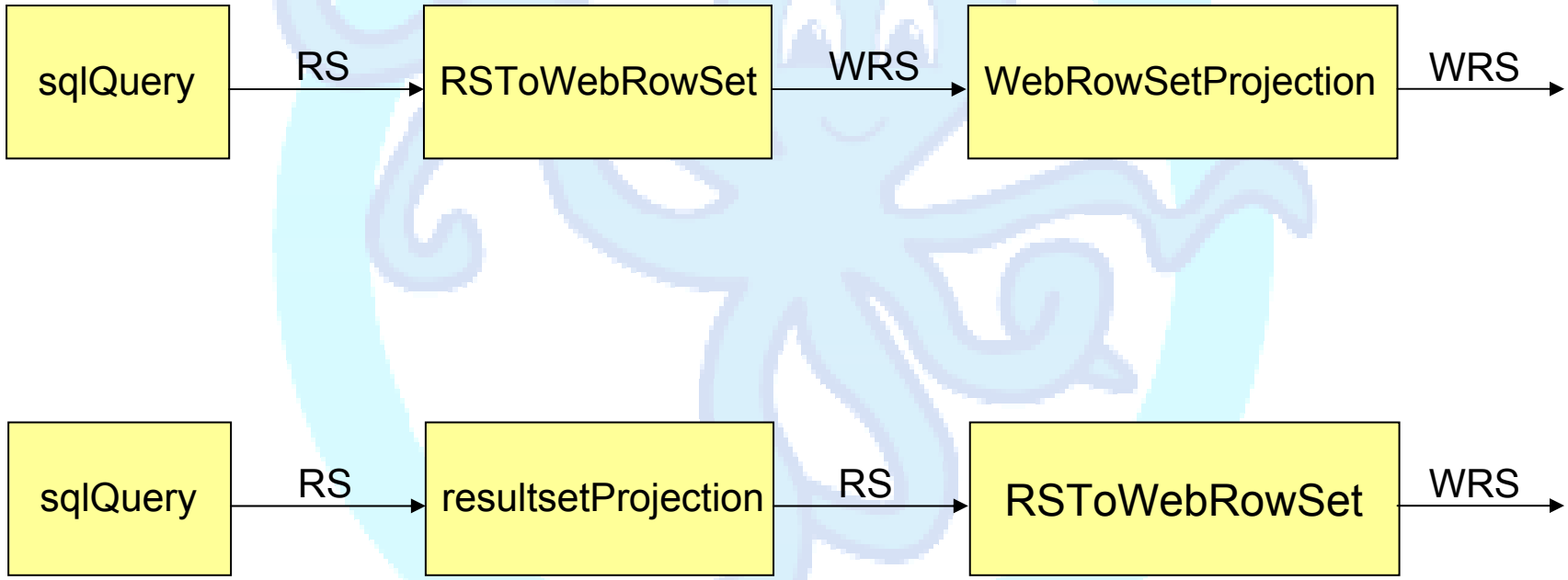
OGSA-DAI

Workflow Reconstruction

- Be able to identify patterns
  - Describe workflow nodes/edges
- Reconstruct part of the workflow graph
  - Performance
- Some example scenarios

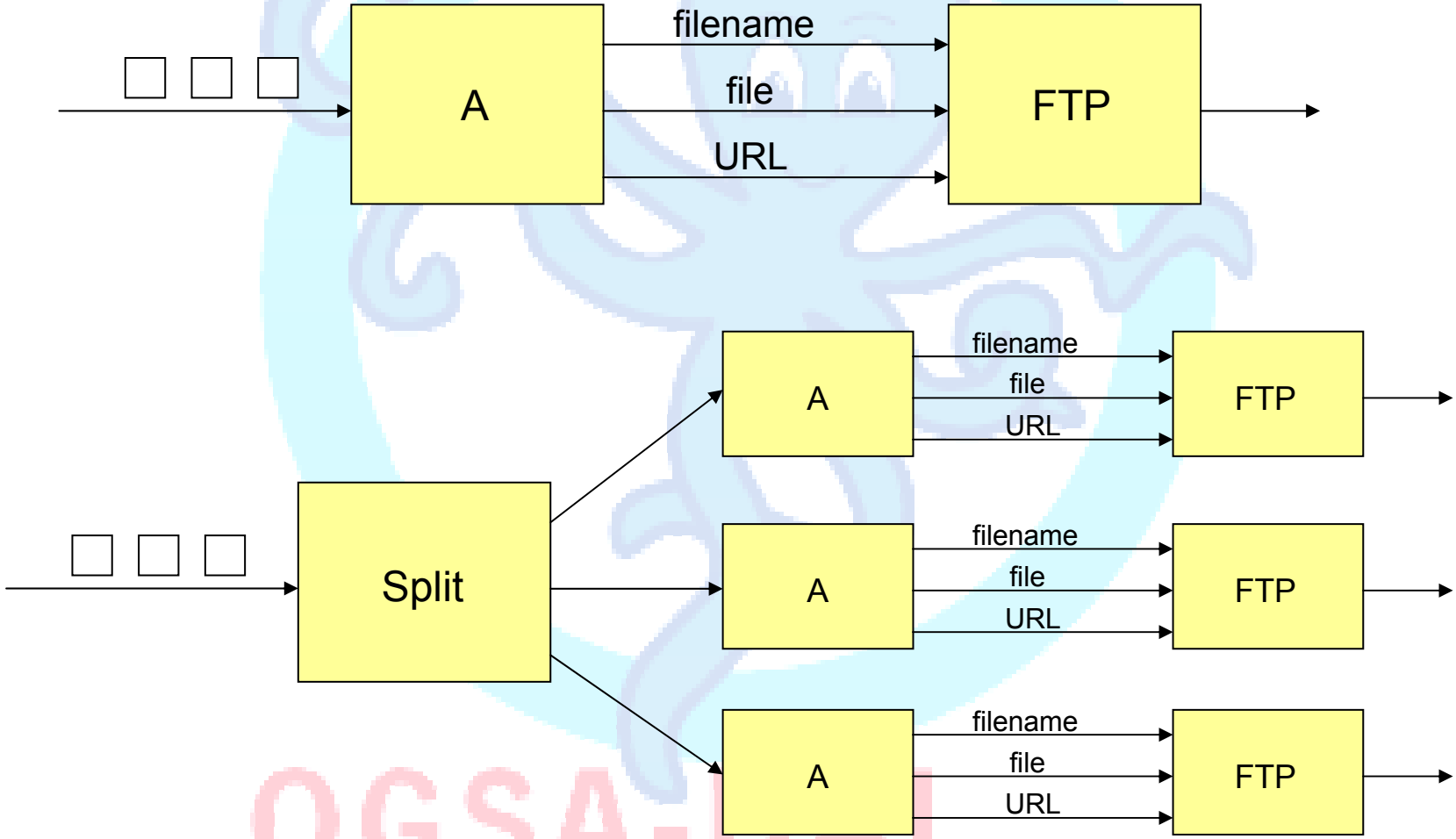


Scenario (1)



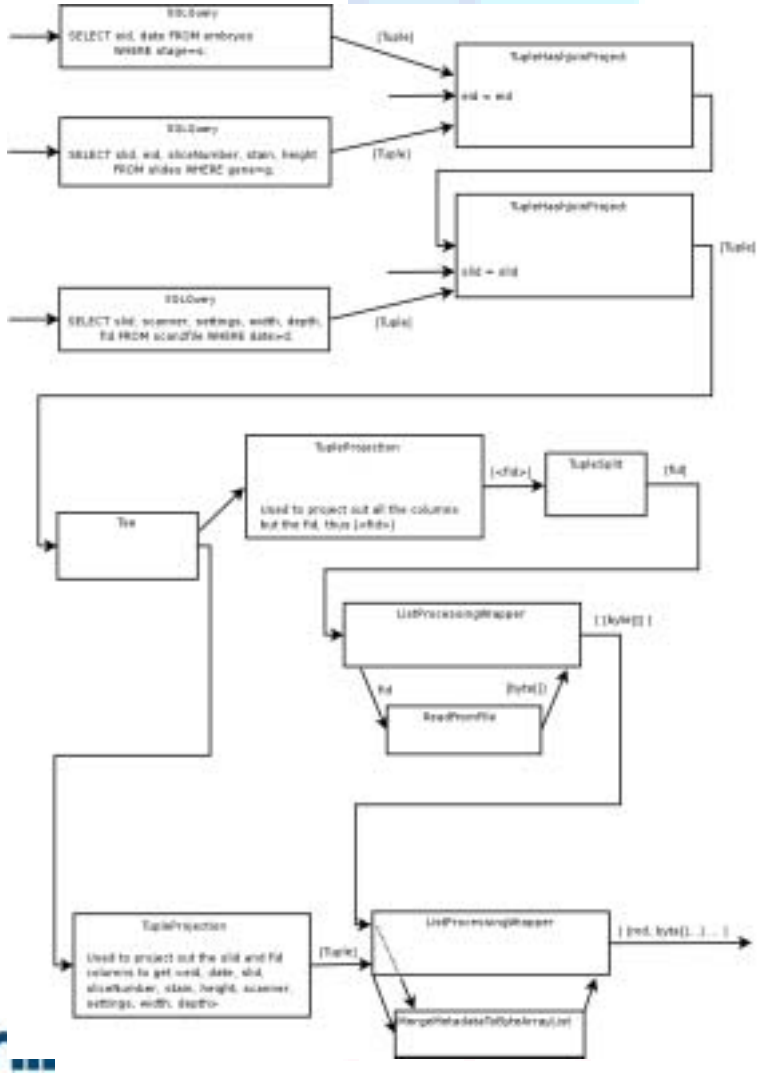
OGSA-DAI

Scenario (2)



OGSA-DAI

OGSA-DAI workflows



- Data-driven workflows to solve problems
  - –Obtain scan data for scans since date **d** of embryos in stage **s** showing expression of gene **g**.

Overview

- What is OGSA-DAI
- What you can do with OGSA-DAI
- How do you use OGSA-DAI on the NGS
- Workflow in OGSA-DAI
- What's coming up in OGSA-DAI v3.0?
- Where you can get more information

OGSA-DAI 3.0

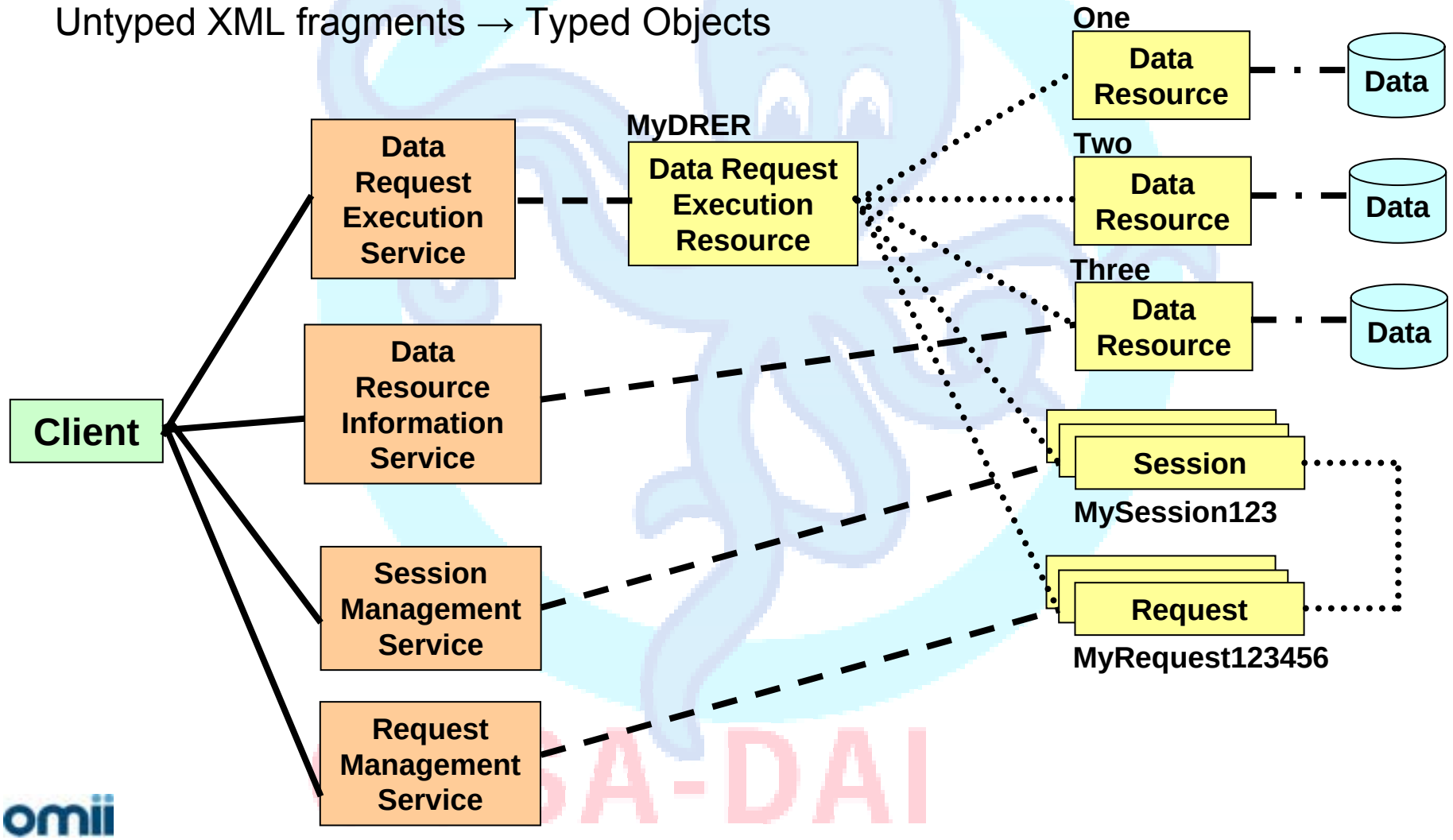
- Top to bottom rewrite
- New service and resource model
- APIs to write new web service layers
- Persistence module
- New activity framework
  - new input and output types
  - invocation
  - iteration
- New security framework
- Released Q2 2007

OD3: What does this mean?

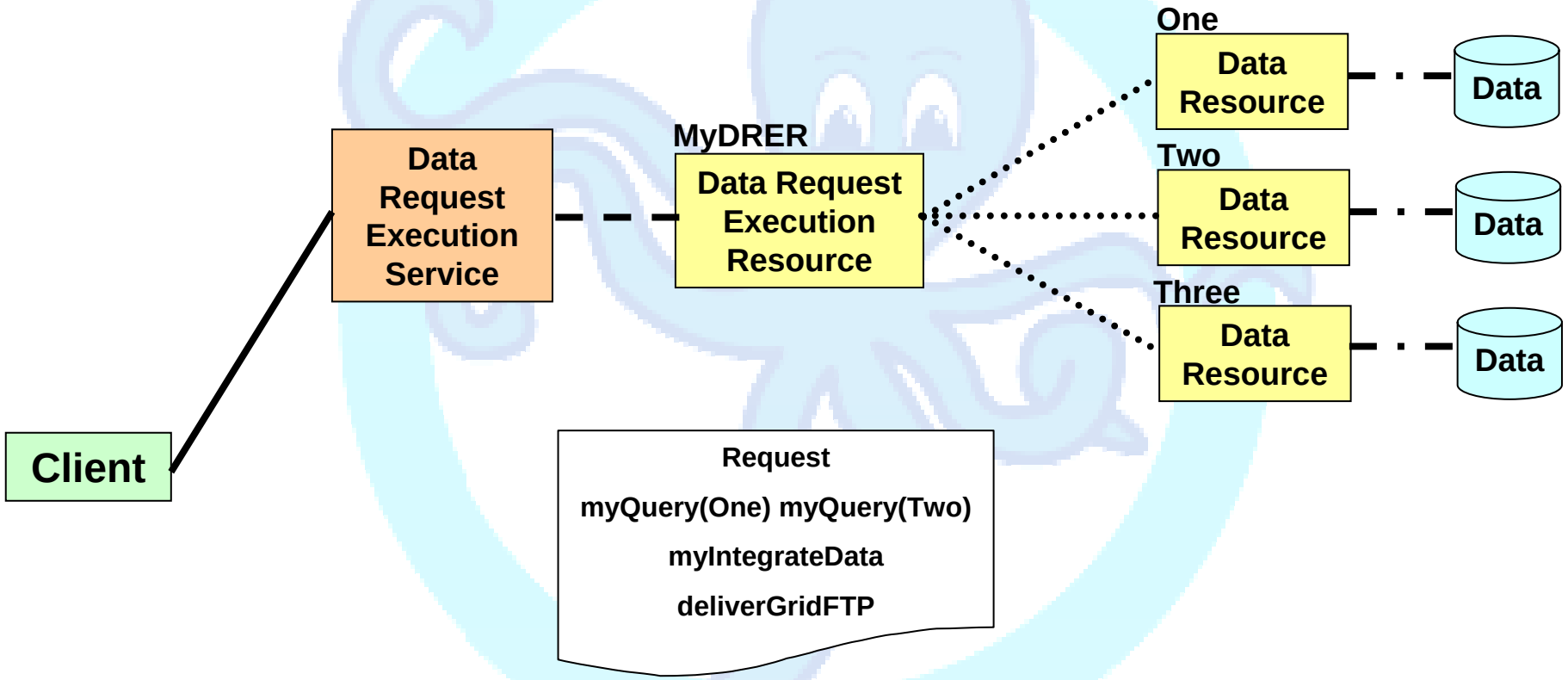
- You can:
  - Chain OGSA-DAI services together to create powerful data-driven workflows.
  - Create workflows that integrate and transform data from multiple data resources, including accessing multiple data resources from within the scope of a single OGSA-DAI request.
  - "Reskin" OGSA-DAI with application-specific presentation layers to fit particular domains (e.g. DAIS, OGC, etc).
    - also means it's easier to maintain different flavours:
      - OMII-UK, GT4, UNICORE GS, GRIA, gLite etc.
  - Develop application-specific activities easily and without resorting to XML manipulation.

An expanded resource model

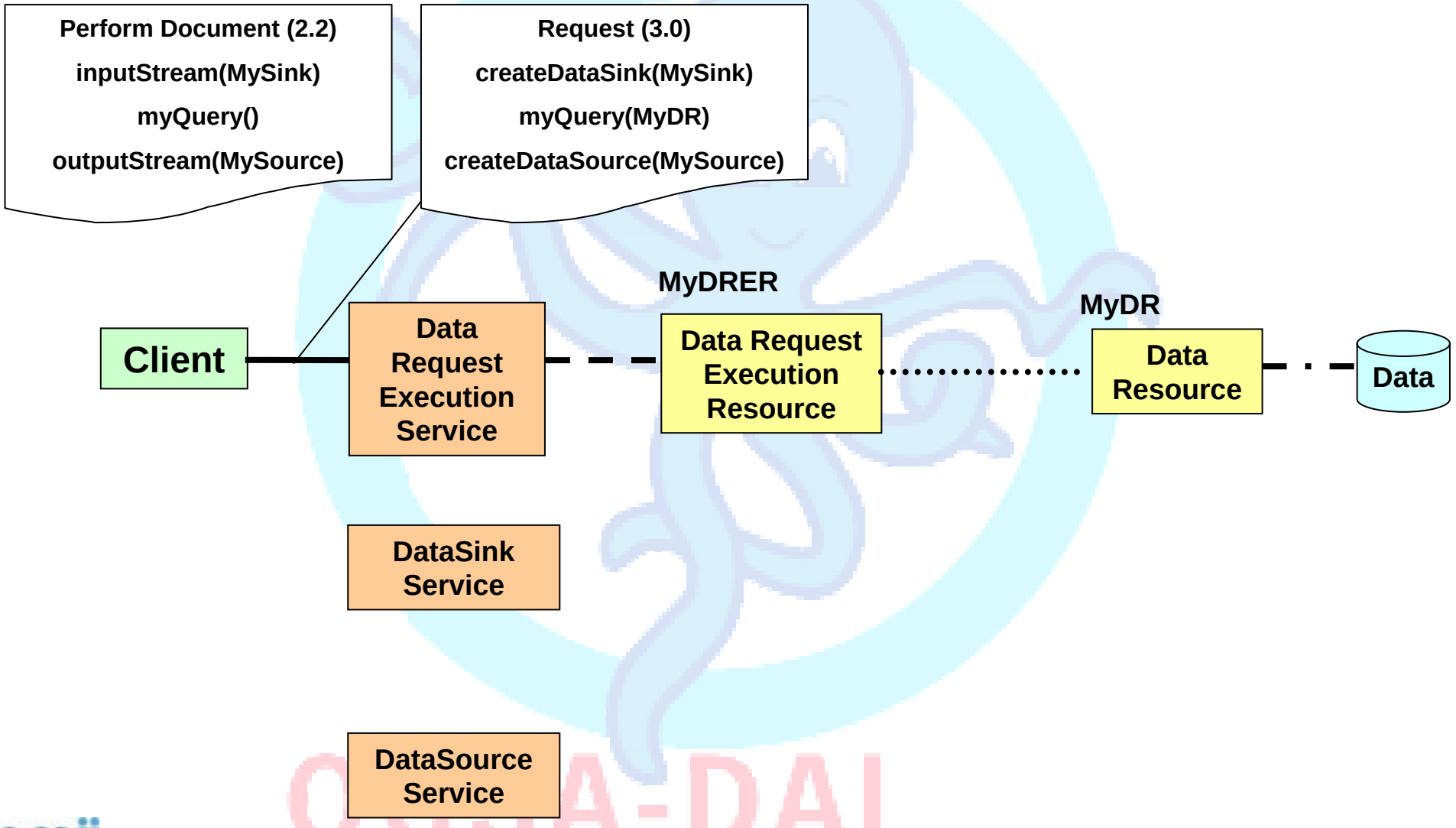
Perform Document → Request
Response Document → Request Status
Untyped XML fragments → Typed Objects



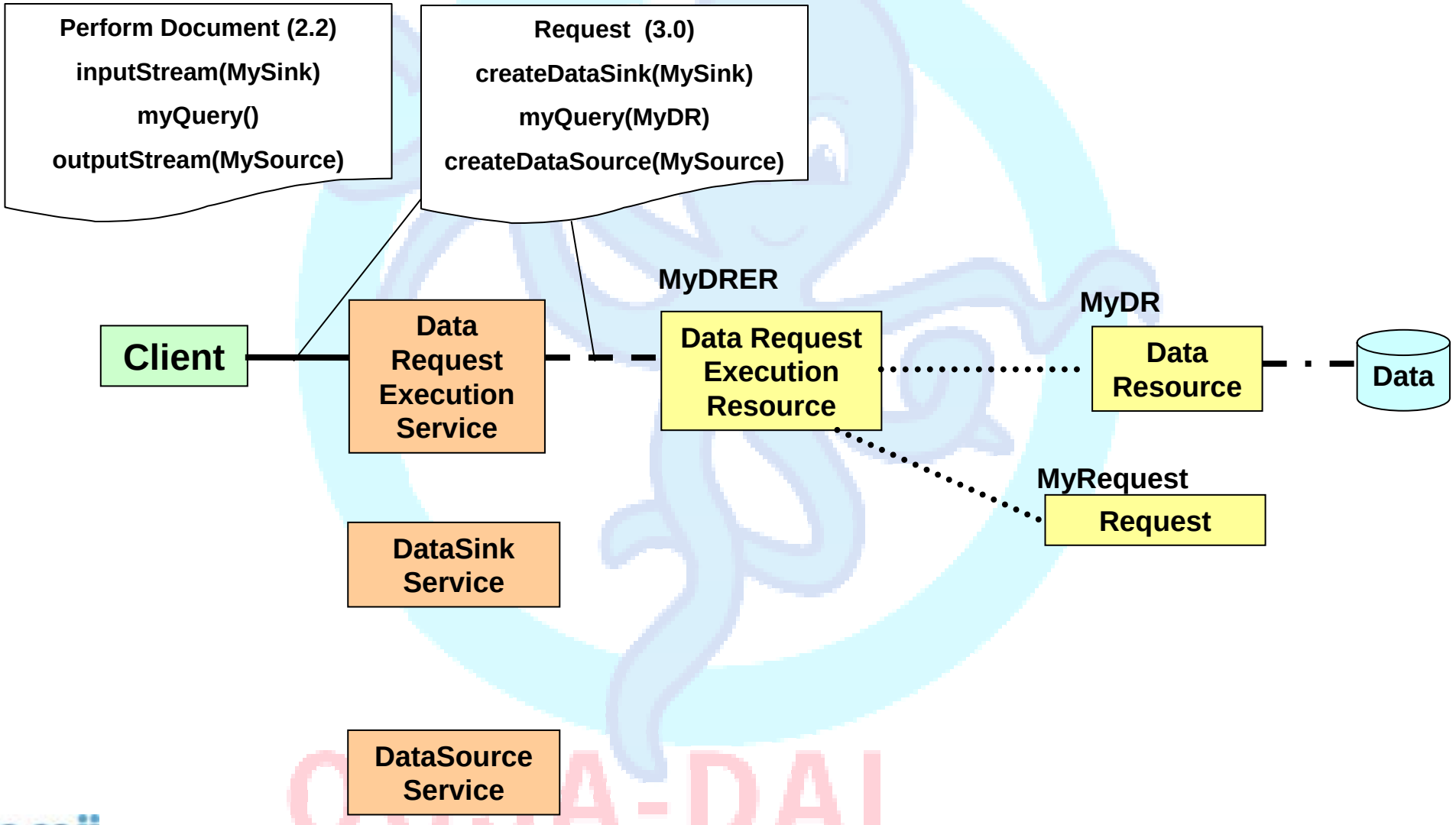
Multi-resource requests



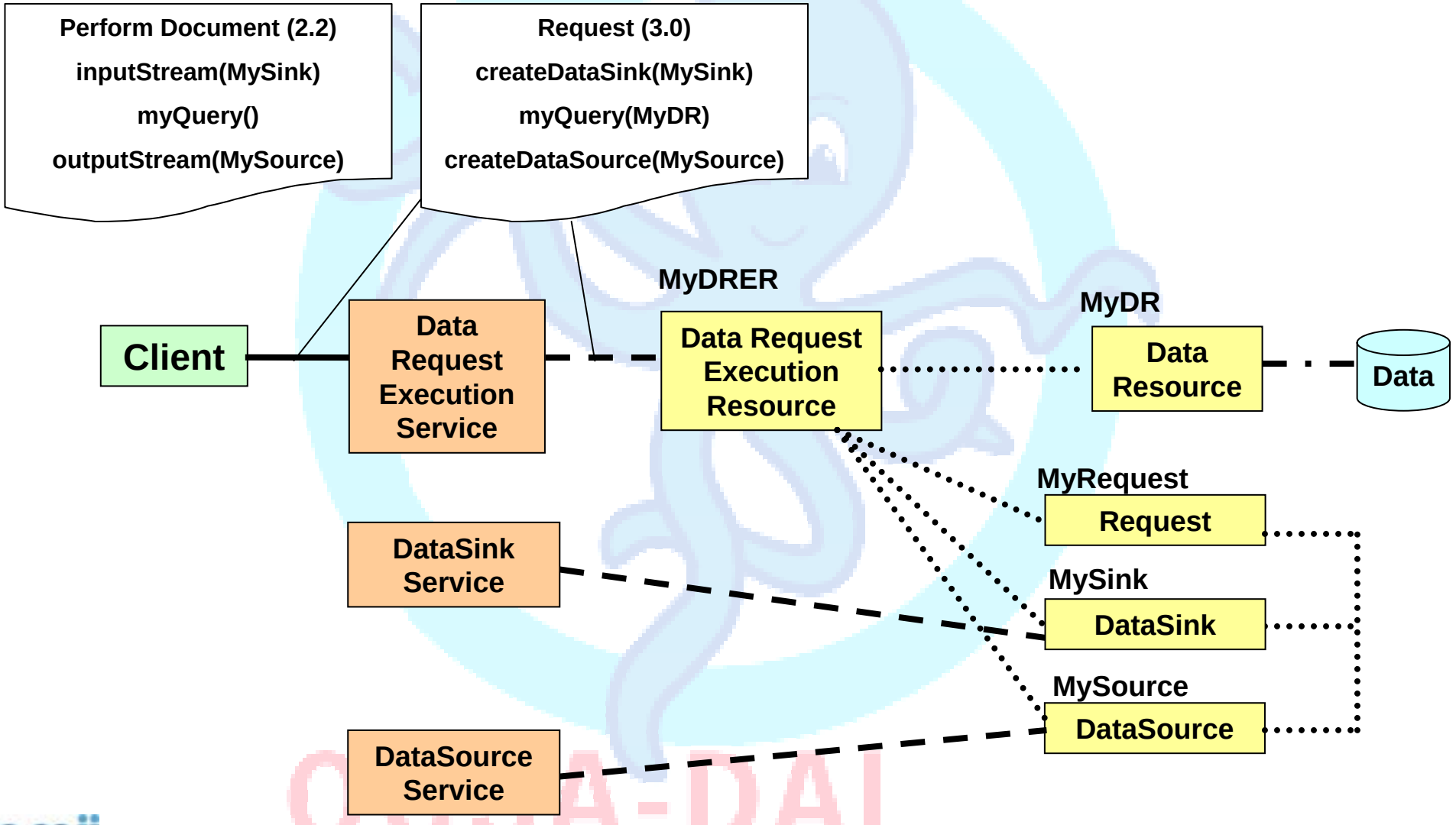
Data transport



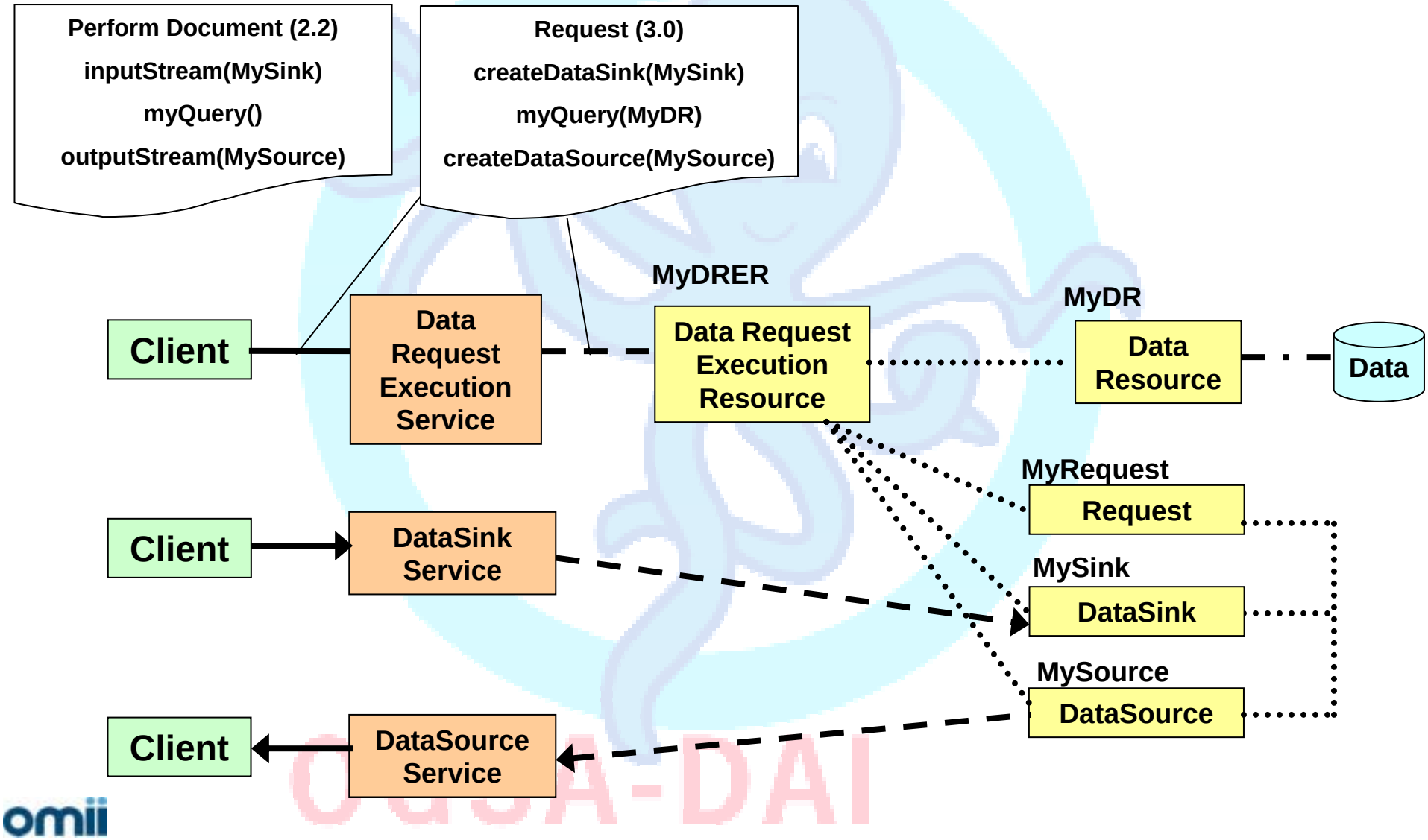
Data transport



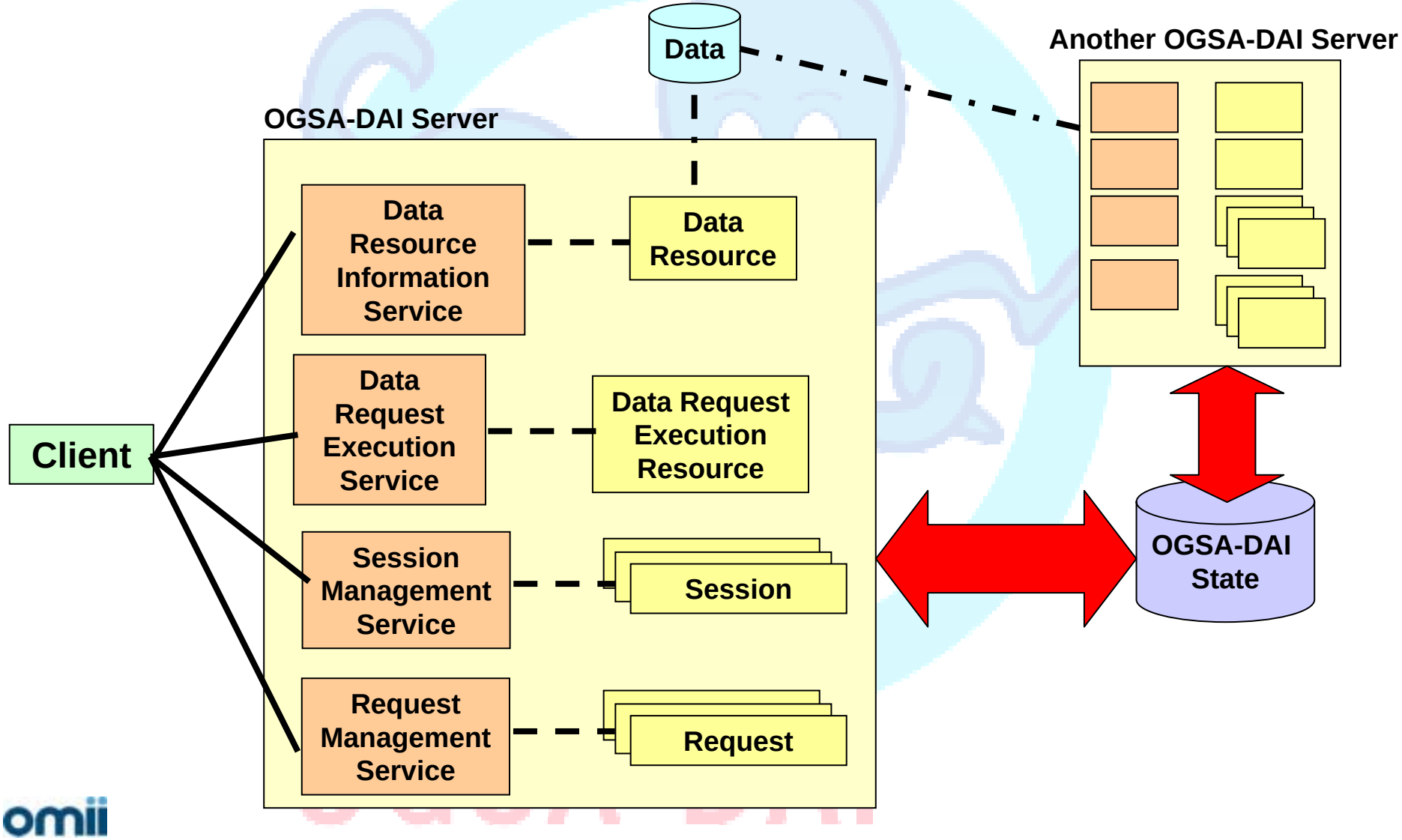
Data transport



Data transport



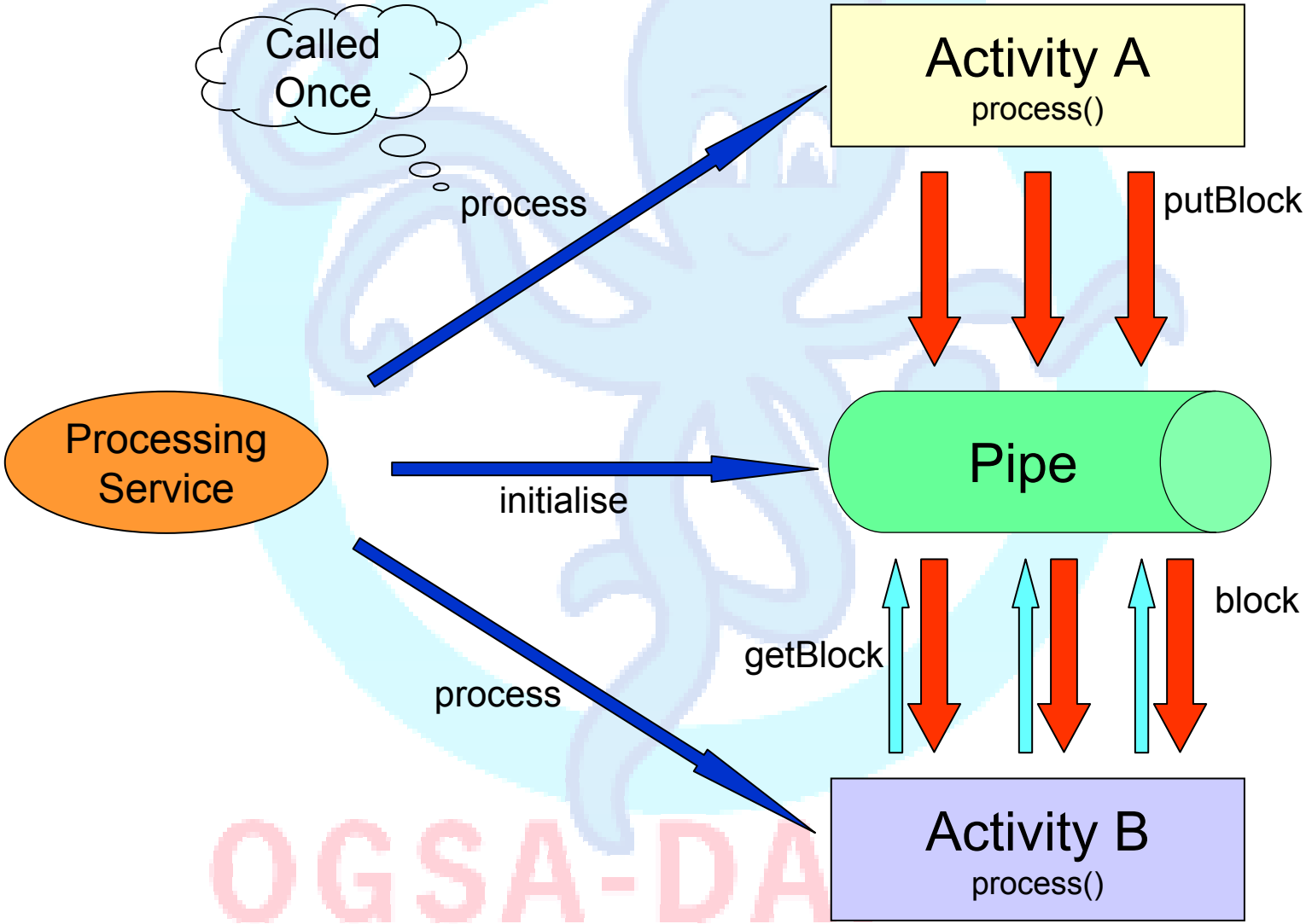
Persistence



Persistence

- Caching
  - Reduce overheads of OGSA-DAI-database communications
- Allows configuration and state to persist between container shutdowns/crashes
- Sticky resources
  - e.g. data sources and sinks
  - In-memory – only available via server that created them
- API
  - File-based implementation for backwards compatibility
  - Relational implementation for compatibility with common OGSA-DAI databases
  - Extensibility point

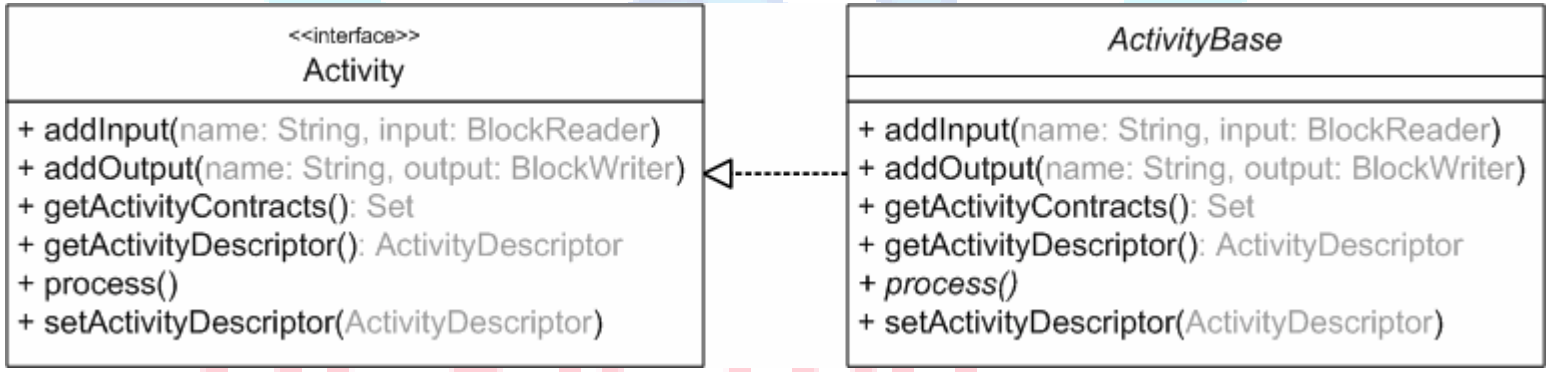
New Activity Model



OGSA-DA

Activity Interface

- All activities must implement the *Activity* interface
- Activities may extend the abstract *BaseActivity* class which implements common functionality => only need to implement **process()**



Activity Extension Interfaces

- New interfaces extend the base activity to provide access to necessary information
  - Activities that access data resources
  - Activites that create new requests
  - Activities that are configurable
  - …

Monitoring

- Monitoring Framework: Provides *Listener* interfaces
- *Activity Listeners*: Receive activity events, for example
  - Activity initialised
  - Activity processing
  - Error occurred
  - Processing completed
  - Custom events …

OGSA-DAI

More Listeners

- *Pipe Listeners*: Receive pipe events, for example
  - Block produced
  - Block consumed
- Various listeners may be implemented to support logging, auditing, debugging etc.

OGSA-DAI

New security framework

- Information collected at various points in the system
- Policy Decision Points similar to previous OGSA-DAI version
- Allows extensibility e.g. VOMS
- Tries to authorize at highest sensible level
  - but if required will call back to higher layers from sub request object level
- *I am not a security expert*
- Developed in conjunction with SIMDAT and inteligrid projects

OGSA-DAI

Summary

- OGSA-DAI is middleware which allows uniform access to data sources which are:
  - diverse
  - heterogeneous
  - independently curated
- It is designed to be:
  - efficient
  - extensible
  - portable
  - easy to develop
- It brings together remote data sources at run-time.
  - and reduces round trips through use of workflows

Further information

- See more projects using OGSA-DAI:
  - http://www.ogsadai.org.uk/about/projects.php
- And what they've been doing:
  - http://www.ogsadai.org.uk/about/success_stories/
- Learn to program OGSA-DAI:
  - http://www.ogsadai.org.uk/documentation/ogsadai-wsrf-2.2/doc/clients/clienttoolkit/index.html
- See what's coming up in OGSA_DAI 3.0:
  - http://www.ogsadai.org.uk/documentation/Design_documents/

Further information

- The OGSA-DAI Project Site:
  - http://www.ogsadai.org.uk
- The DAIS-WG site:
  - http://forge.gridforum.org/projects/dais-wg/
- OGSA-DAI Users Mailing list
  - users@ogsadai.org.uk
- Formal support for OGSA-DAI releases
  - http://bugs.ogsadai.org.uk
- OGSA-DAI training courses (live and online)

Neil Chue Hong
EPCC
N.ChueHong@epcc.ed.ac.uk
+44 131 650 5957