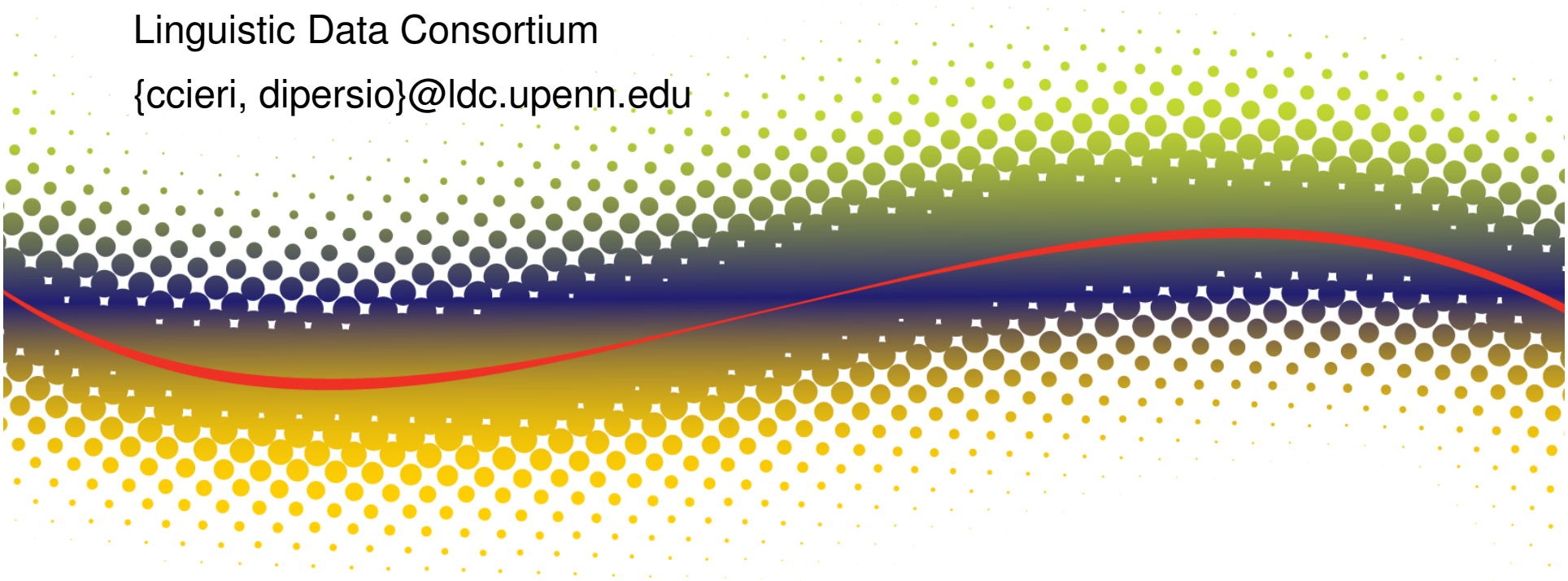
A License Scheme for a Global Federated Language Service Infrastructure

Christopher Cieri, Denise DiPersio

Linguistic Data Consortium

{ccieri, dipersio}@ldc.upenn.edu

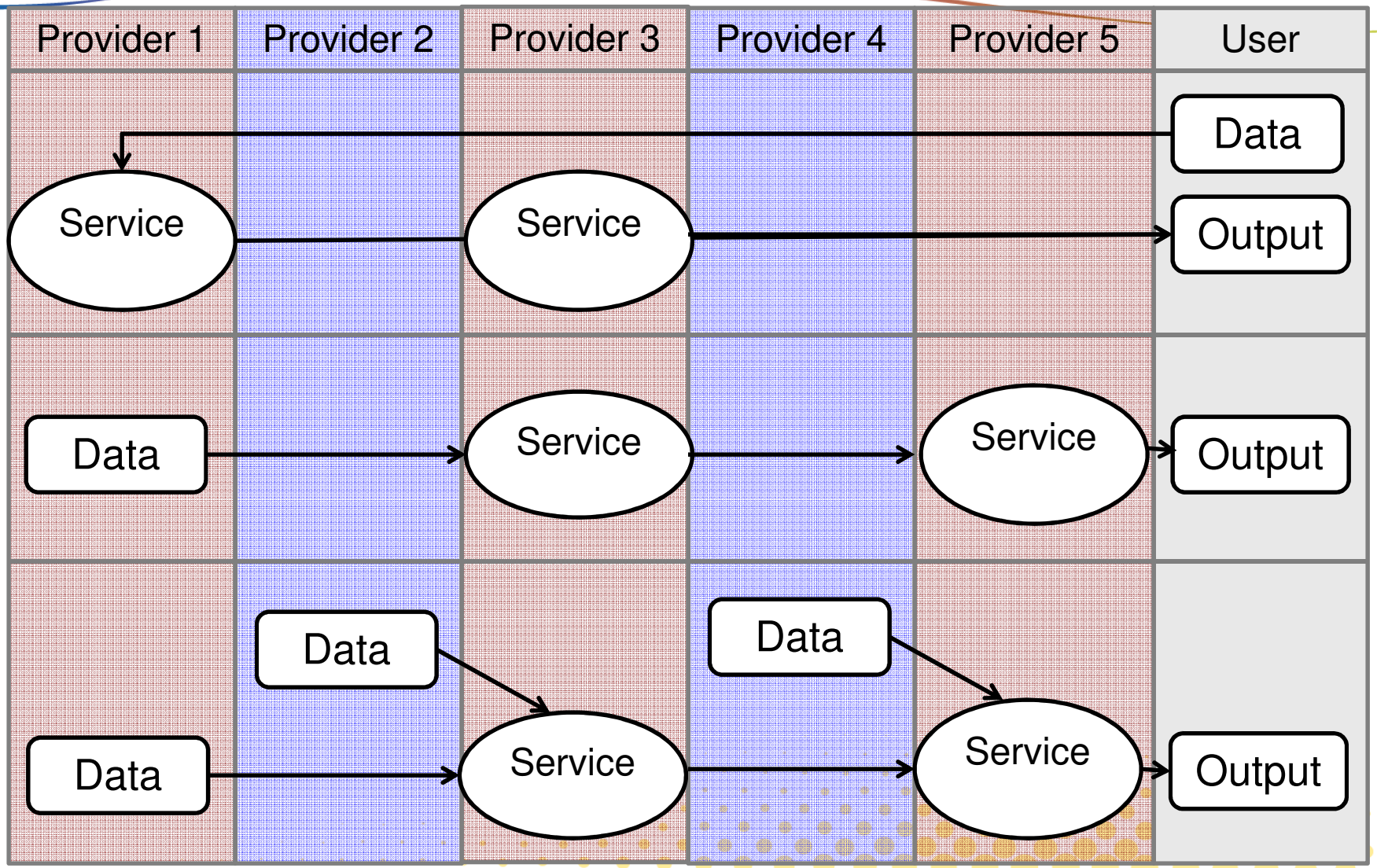


Licensing in the Grid Environment

- Tension between traditional resource licensing and grid licensing
  - Data centers and large-scale language resource distribution came of age in the early 1990s before a developed Internet, increased computing power, the cloud
  - Early goal: to convince real-world data providers that language researchers, human language technology developers use their data for a benign purpose (e.g., not competitive) and to further scientific progress
  - Agreements cover distribution from one point to another (data center → user)
    - User cannot redistribute data set to others
- Result that language resources typically controlled within a small community
- Language service infrastructures challenge that model
  - Resources live on, and are processed over, the web
  - Potentially many users – the virtual community
  - Processing steps may involve several platforms, tools
  - Licenses/permissions must be handled on the fly

Linguistic Data Consortium

Sample Pipelines



Grid Complexities

- Multiple stakeholders
  - Grid operators (repository host; maintains software, services)
  - Service providers (access to data, software)
  - Users
- Numerous configurations of web services, software repositories, clouds, clusters, local processing
- Distributed responsibility: no single operator, provider or user controls the whole
  - Magnified in federated grids
- Disparate approaches to managing behavior including intellectual property rights
  - control providers and/or
  - control users
  - laissez faire

License Variables and Gaps

- License types differ
  - Software licenses: typically regulate use of software and derived works (other software)
    - No software licenses surveyed restrict software output
  - Data licenses: typically control use of data and derived works (other data, which is output)
- No calculus for conflicting, ambiguous terms
  - 1 LR prohibits commercial use; 1 LR allows it. What is the impact on processed output that uses both?
  - Derivative work – transformative work continuum
- Grid licensing not so different from established practice
  - Terms apply to future conduct; community's established good behavior
  - BUT fluid grid environment means users might overlook license; planning required to avoid this

Extant Grid License Schemes 1/2

- The Language Grid (NICT, Kyoto University)
  - Providers designate use categories – non-profit, research, commercial
  - Workflow composer displays relevant licenses; permitted uses shown when browsing services
- META-SHARE
  - Infrastructure to promote resource description and sharing
  - New language processing layer
  - License types: CC, META-SHARE CC, No Redistribution
    - Constraint-free within network to the extent possible
- PANACEA
  - Provides chained web services for data processing and crawls the web to develop data sets on demand for research use
  - Resource providers clear rights to data and tools they contribute to PANACEA whether they own them or not

Extant Grid License Schemes 2/2

- LinguaGrid
  - Open to different operators with configurable service access policies; research and commercial licensing
  - Built on Language Grid infrastructure
- CLARIN
  - Networked federation of European data repositories and service centers accessible by network participants
  - Range of license options (including CC) with ability to carve out user groups (e.g., META-SHARE members)
- LAPPS Grid
  - Accommodates range of license types; pipelines should succeed
  - License constraints accumulate as pipeline is constructed
  - A few constraints are required; pipeline is blocked if conditions not met
  - Most limitations are presented as notifications that users must acknowledge before workflow commences

Identifying License Dimensions

- Language Resources
  - *Owned* by: user, public domain, copyrighted, otherwise constrained (contract)
  - Limits on *use, sharing*: non-commercial, no derivative works, share-alike, attribution, no distribution, other miscellaneous
  - Limits on *users*: non-commercial organizations
- User
  - Not licensed or licensed
    - To group defined by enumeration – LDC user database
    - To group defined by features: academic, non-academic NFP, government, (pre-)commercial
- Use: education, basic and applied research, (commercial) technology development, technology evaluation, technology deployment, resale
- Services: permitted to improve training from input language resources
- Processing
  - Creating a derivative work: transcription
  - Creating a transformative work: untied language model, frequency list
- Some combinations are in conflict

Linguistic Data Consortium

LAPPS Grid Software

License	Software
Apache 2.0	Language Grid software, NLTK, ANC2G0, UIMA, OAQA, Uimafit, guava-libraries, ActiveMQ, AnyObject, Jaxws-maven-plug-in, Jetty, OpenNLP
BSD	Hamcrest, NERsuite, CRFsuite (in NERsuite)
CDDL 1.1	Jaxws-rt
CPL 1.0	MALLET, AGTK, JUnit
Eclipse 1.0	logback (v1.0), Jetty
HTK-Cambridge	HTK
MIT	Mockito, libLBFGS (in NERsuite), GIZA (v3)
Python	NLTK
Wordnet	Genia tagger library (in NERsuite)

Linguistic Data Consortium

License Types and Constraints 1/2

License	Redistribution	Use	Derivative Use	Attribution	Share Alike	Fee
CC-Zero	Yes	Commercial	Commercial	No	No	No
CC-BY	Yes	Commercial	Commercial	Yes	No	No
CC-BY-SA	Yes	Commercial	Commercial	Yes	Yes	No
CC-BY-ND	Yes	Commercial	None	Yes	No	No
CC-BY-NC	Yes	Research	Research	Yes	No	No
CC-BY-NC-SA	Yes	Research	None	Yes	Yes	No
CC-BY-NC-ND	Yes	Research	None	Yes	No	No
GPL (v 2,3)	Yes	Commercial	Commercial	Yes	Yes	No

Acknowledgment:
Meta-Net created an earlier version of this table that we extend

Linguistic Data Consortium

License Types and Constraints 2/2

License	Redistribution	Use	Derivative Use	Attribution	Share Alike	Fee
Apache 2.0	Yes	Commercial	Commercial	Yes	No	No
BSD	Yes	Commercial	Commercial	No	No	No
CDDL 1.1	Yes	Commercial	Commercial	Yes	Yes	No
CPL 1.0	Yes	Commercial	Commercial	No	No	No
Eclipse 1.0	Yes	Commercial	Commercial	Yes	Yes	No
HTK-Cambridge	No	Commercial	Commercial	No	No	No
MIT	Yes	Commercial	Commercial	No	No	Yes
Python	Yes	Commercial	Commercial	Yes	No	No
WordNet	Yes	Commercial	Commercial	Yes	No	No
LDC FP Member	No	Commercial	Commercial	No	No	No
LDC NFP Member	No	Research	Research	No	No	No
LDC Non-member	No	Research	Research	No	No	Yes

Linguistic Data Consortium

The Essential Constraints

Constraint	Values
Redistribution	Yes/No
Use	Commercial/Research Only
Derivative Use	Commercial/Research Only/None
Transformative Use	Commercial/Research Only /None
Attribution	Yes/No
Share Alike	Yes/No
Fee	Yes/No
Other Specific License, Constraint	--

Linguistic Data Consortium

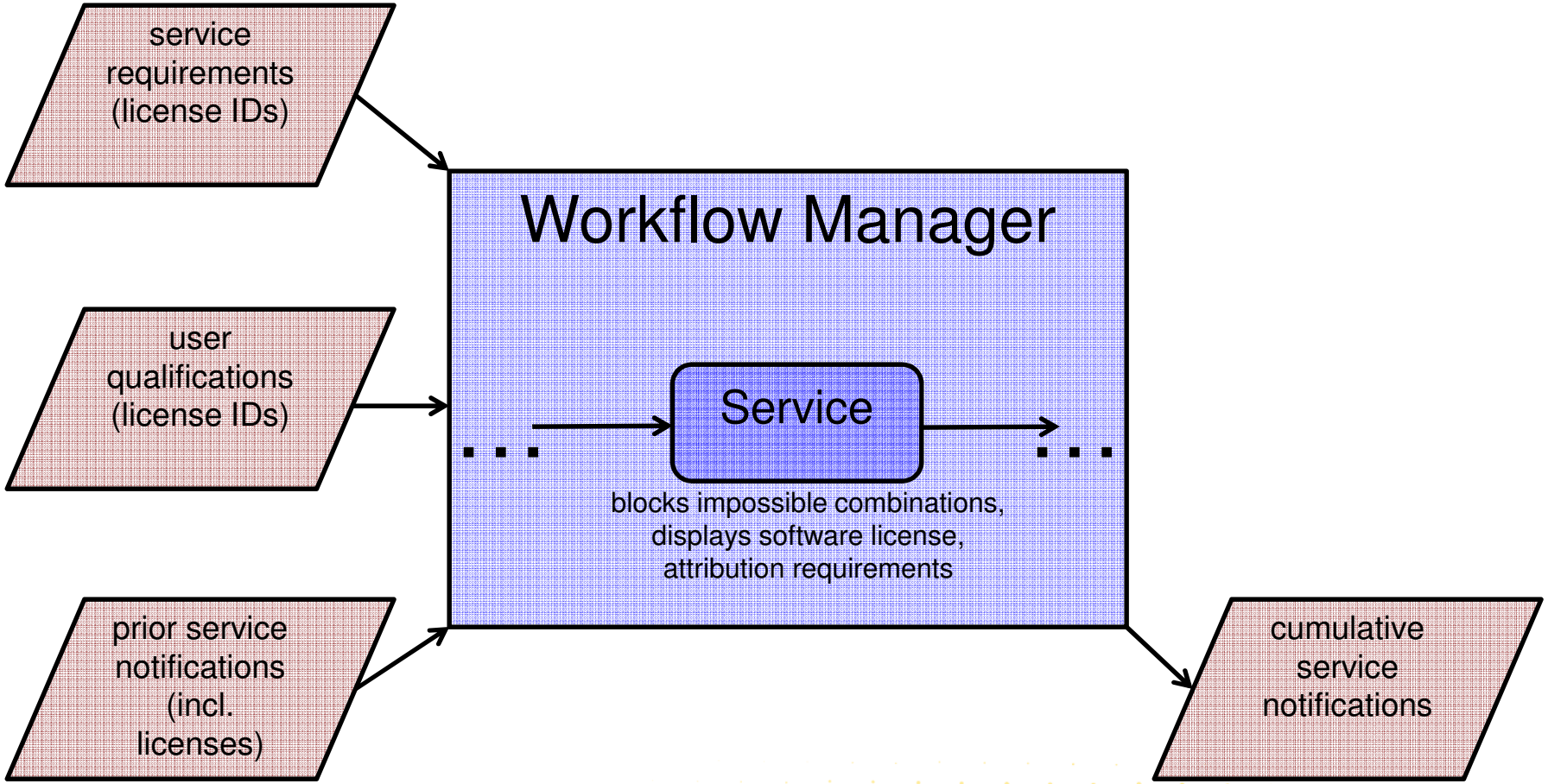
Enforcing Constraints in the Workflow

Constraint	Action
Redistribution	Notify
Use	Notify
Derivatives Use	Notify
Attribution	Notify
Share Alike	Notify
Fee	Require
Other Specific License	Require
Other Specific Constraint	?

License Model for Federated Grid

- User authenticates to initiate session; requests services from workflow manager
- Workflow manager requests license conditions from each requested resource, service
  - May query API, data center regarding user's status and satisfaction of license conditions
- Manager accumulates summary of click-through licenses
- Manager presents license summary with links to text
- User clicks to agree to all terms before pipeline will start
- Few combinations are barred
- Most combinations appear as notifications

Federated Workflow Based on LAPPS Approach



Issues for Federation Licensing and Operation

- Should there be a federation agreement?
  - Effect of pre-existing grid agreements
- Can authorized users on one grid access all grids in the federation without extra steps?
- Interaction between grid operators, service providers
  - Offering commercial services on a research grid?
- Who is responsible for maintaining grid operation/uptime?
- What are the consequences for disrupting the grid?
- Should operators, providers provide warranties for resources and services?
- What is the approach to dispute resolution?
- What user information can be collected/used by the federation?
- How are providers/users notified about changes to grid services?

Linguistic Data Consortium

Conclusion

- Traditional license management does not translate well to language service infrastructures
- Solution must account for workflows combining multiple resources with various license terms that are executed on the fly
- Workflow managers can mediate licenses allowing credentialed users to construct complex pipelines and execute them
- Applying that model to an open global federated infrastructure requires consideration of administrative and operational issues
- We look forward to working with the community to achieve successful federation