

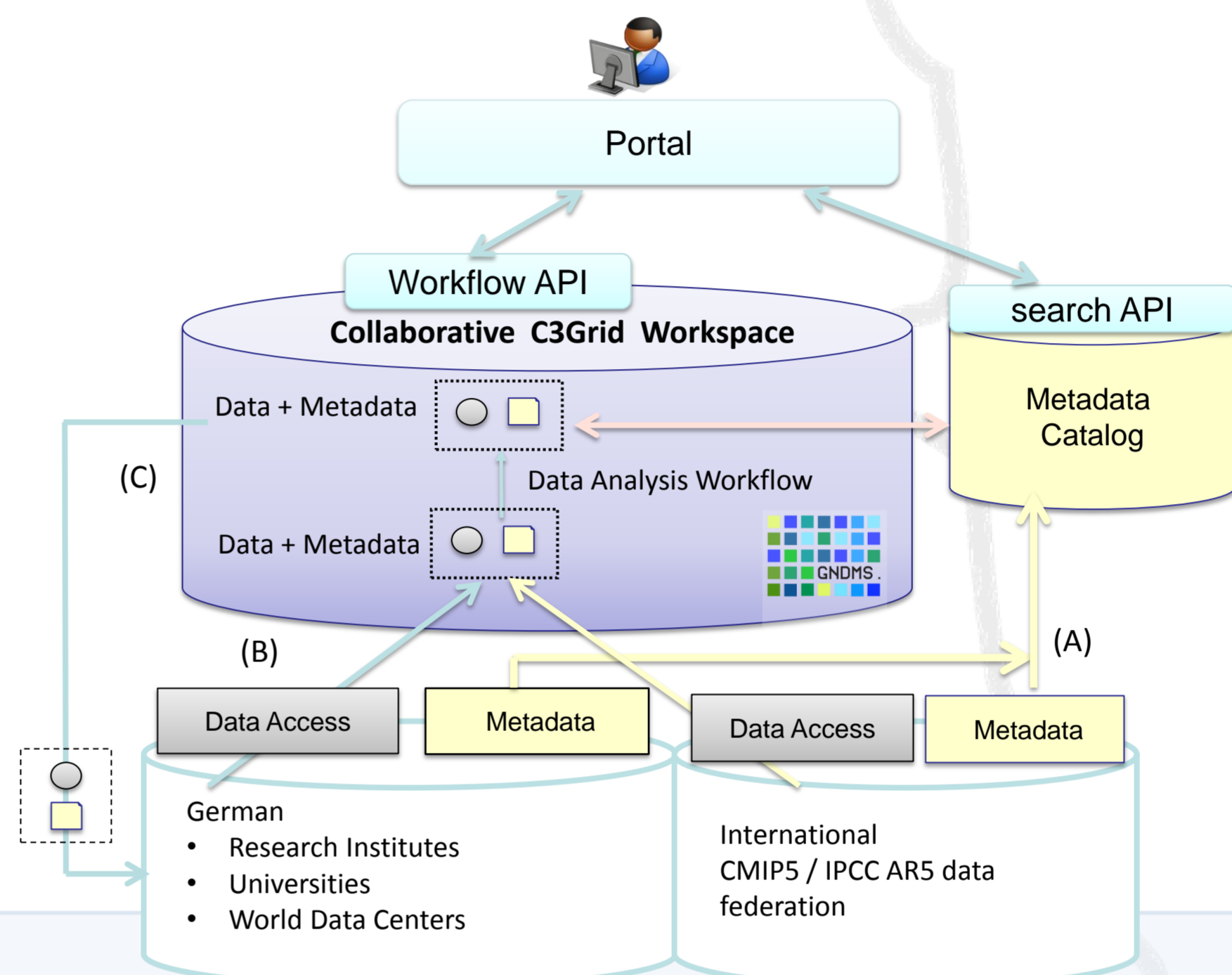
Motivation

- Aufgrund explosionsartig wachsender Klimadatenbestände ist eine Dateninfrastruktur notwendig, die Klimaforscher bei komplexen Datenauswertungsaufgaben unterstützt.
- Der bisherige Arbeitsablauf von Klimaforschern war (grob skizziert):
 - (1) Suche nach geeigneten Archiven/Datensätzen, (2) Download aus Archiven mit teilweise sehr unterschiedlichen Zugriffsmethoden und (3) Prozessierung / Analyse der Daten „zu Hause“
- Dieser Ansatz ist angesichts weltweit verteilter Klimadatenarchive mit Datenbeständen im Petabyte-Bereich nicht mehr gangbar.
- Die C3Grid Infrastruktur stellt deshalb einen verteilten virtuellen Kollaborationsbereich zur Verfügung, der
 - (1) eine einheitliche Suche in angeschlossenen Datenarchiven erlaubt und (2) einen einheitlichen Datenzugriff bereitstellt, sowie (3) eine verteilte kollaborative Datenverwaltung und Datenprozessierung unterstützt.

Datenintegration

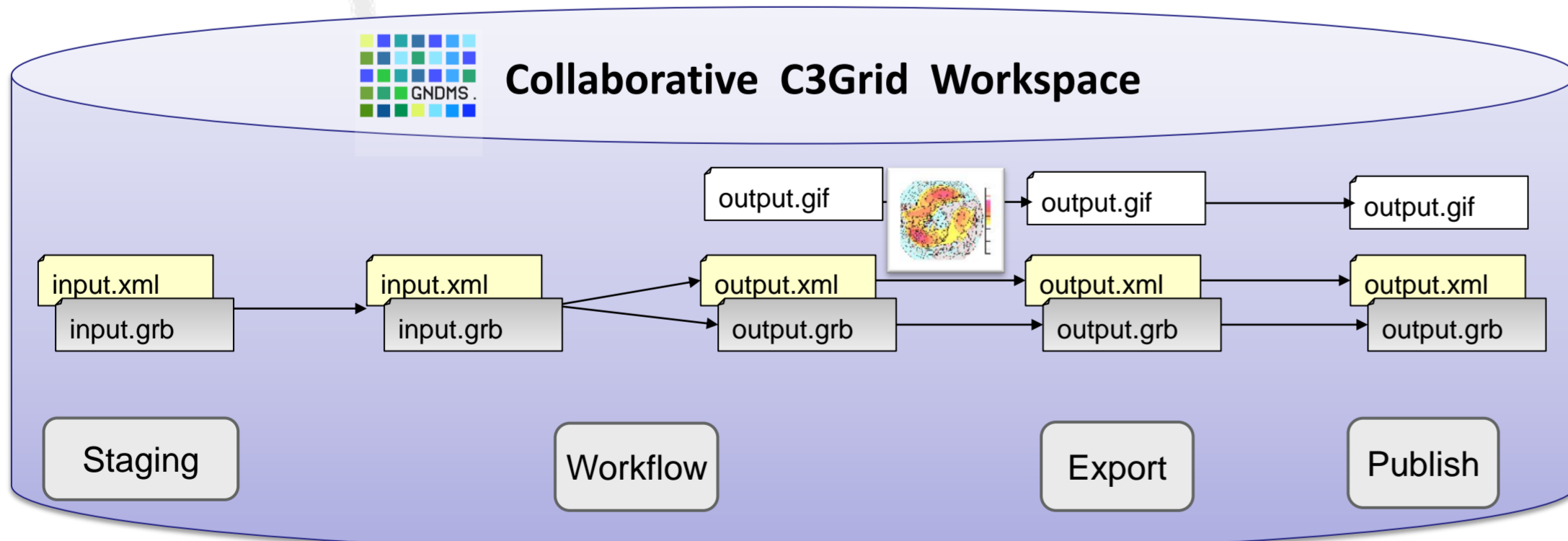
Klimadatenanbieter werden über standardisierte Schnittstellen (REST, HTTP, GridFTP) in C3Grid integriert:

- Nationale Archive implementieren C3Grid Daten- und Metadatenchnittstellen.
- C3Grid unterstützt die Schnittstellen der internationalen ESGF Datenföderation.
- Durch C3Grid werden Metadaten in einem zentralisierten Metadatenkatalog gesammelt und sind über ein gemeinsames Suchinterface auffindbar.
- Daten werden mit zugehörigen Metadaten im C3Grid Workspace verwaltet; Ergebnis-Daten können extern zugreifbar gemacht werden.



Kollaborativer C3-Workspace (GNDMS)

- Temporärer Grid-Zwischenspeicher für Daten
 - unterstützt die C3Grid Security-Infrastruktur und Zertifikatsdelegation
 - Datentransfer (Im-/Export) über gridFTP und HTTP
 - räumt sich selbst wieder auf (Daten sind mit Lebensdauer annotiert)
 - trennt verschiedene Datensätze (ggf. mehrere Dateien) logisch voneinander (sog. Slices und Slice-IDs)
 - Workflows können Slice-IDs als Eingabedaten verwenden, um z.B. Ergebnisdaten früherer Workflows weiter zu analysieren
 - Anbindung an Workflow-Scheduler (Abschätzung von Daten-Staging-Zeiten etc.)
- C3-Workflows
 - erzeugen stets neue, für das Ergebnis gültige, ISO-Metadaten mit Provenienz-Information
 - Ergebnisse sind so nachvollziehbar und reproduzierbar und selbst wieder gleichwertige Eingabe-Datensätze im C3Grid
 - Workflows modularisieren die Datenanalysefunktionalität



Data Provider



Compute Provider



Portal

Ausblick

Gemeinsames Arbeiten an Ergebnisdaten:
Nutzer kann Workflow-Ergebnisse freigeben

- einer Gruppe (mehrere Grid-Nutzer)
- öffentlich freigeben

C3Grid Datenmanagement GNDMS erlaubt Publishing:

- Ergebnisse liegen auf der Export-Site
- Transfer auf sog. Publishing-Site (mit Metadaten) (weitere GNDMS-Instanz)
- Publishing-Site liefert Metadaten an Metadatenkatalog
- Publishing-Site stellt Schnittstelle für Datenstaging bereit → Ergebnisdaten sind in C3Grid such- und nutzbar



Weiterführende Informationen unter

- www.c3grid.de
- esgf.org
- gndms.zib.de

