JISC

PREPARDE

D2.1 Journal workflows from author submission of datasets and papers, through review to publication.

D2.2 Data repository workflows from ingestion of data, through data centre technical review, to DOI assignment to dataset

Project Information			
Project Identifier	*To be completed by JISC*		
Project Title	PREPARDE: Peer REview for Publication & Accreditation of Research Data in the Earth sciences		
Project Hashtag	#preparde		
Start Date	1 July 2012	**End Date**	30 June 2013
Lead Institution	University of Leicester		
Project Director	Dr Jonathan Tedds		
Project Manager	Dr Sarah Callaghan		
Contact email	sarah.callaghan@stfc.ac.uk		
Partner Institutions	University of Leicester British Atmospheric Data Centre (BADC) US National Centre for Atmospheric Research (NCAR) California Digital Library (CDL) Digital Curation Centre (DCC) University of Reading Wiley-Blackwell Faculty of 1000 Ltd		
Project Webpage URL	http://proj.badc.rl.ac.uk/preparde/wiki		
Programme Name	*Managing Research Data*		
Programme Manager	Simon Hodson		

Document Information	
Author(s)	Sarah Callaghan
Project Role(s)	Project Manager

Date		Filename	
URL	*If this report is on your project web site*		
Access	☐ Project and JISC internal	☒ General dissemination	

Document History		
Version	**Date**	**Comments**
1	1 March 2013	First draft

Workflows for data publication, from repository to data journal

Introduction

This document holds the workflows captured as part of the PREPARDE project. Workflows were captured for the data centre and journal partners in order to identify points where cross-linking and metadata sharing for data publication would be the most effective.

This report is structured as follows: section 2 lists the workflows received and their source. Section 3 discusses the workflows and attempts to draw some conclusions from them. The captured workflows themselves are then attached to this document.

Workflow listing

Workflows were received for:

- Data Centres
  - CEDA[1] (broken down into type of data submitter)
  - NCAR Earth Observing Laboratory (EOL): Computing, Data, and Software Facility
  - NCAR CISL Research Data Archive (RDA), http://rda.ucar.edu/
  - NERC DOI minting workflow
- Journals
  - Geoscience Data Journal (GDJ)
  - International Journal of Digital Curation IJDC (as a control – representative of workflow for non-data publishers)
  - Dryad (http://wiki.datadryad.org/Submission_System_Workflow)

[1] Centre for Environmental Data Archival – umbrella group containing the British Atmospheric Data Centre (BADC), NERC Earth Observation Data Centre (NEODC) and the UK Solar System Data Centre (UKSSDC)

Discussion

- IJDC workflows are very self-contained, as you'd expect from a non-data paper publisher, where the only flows of communication are internal to IJDC, to the paper author(s) and paper reviewers.
- CEDA workflows vary according to the type of data submitter
    - "engaged submitter" – dataset author is engaged in the process of dataset ingestion into the archive and will answer questions and provide metadata and supporting documentation. Datasets from engaged submitters are most likely to be assigned with DOIs after the ingestion process is completed.
    - "data dumper" – the dataset is provided to the data centre "as-is" with no further supporting information, metadata or contact with the author. In some cases, this is legacy data where the data centre are archiving it to save it from deletion. These datasets are un-likely to be awarded DOIs as they probably do not meet the technical requirements for DOIs. However, if it is determined that these datasets are scientifically important then effort may be found to dig up more metadata/clean up the dataset, and they then might be awarded a DOI.
    - "3rd party data request" – this is when a researcher asks the BADC to broker a transfer of data between them and a 3rd party (e.g. the Met Office). DOIs may or may not be assigned to these datasets, depending on the licensing conditions associated with the transfer of the data between the researcher and 3rd party, and the conditions of storage of the data in the data centre.
- For CEDA (and all the NERC data centres) DOIs get assigned at the end of the ingestion process, and after a few more checks of the dataset to ensure it meets the technical quality requirements to be assigned a DOI.
- Similarly to the CEDA workflows, the NCAR workflows point out that it can take months/years to process and ingest a dataset into the archive.
- Both the NCAR and CEDA workflows have several points where the data submitter is contacted to clarify and/or correct metadata and data.
- Both data centres make a point of notifying their user community when new data is fully ingested and becomes available.
- An obvious cross-linking point into GDJ is in the box "XML info (provided by data centre, via author) is used to populate the data set tagging within the article". It is possible to do this without the input of the author, by simply using the DOI to connect to the dataset's DataCite metadata record (via http://search.datacite.org)and ingesting the appropriate metadata for the data paper that way. More metadata could be scraped directly from the data repository's metadata record, but that would mean that the metadata collected by the data centre needs to be easily mapped to the metadata needed by the journal. An intermediate step would be to automatically ingest the metadata from the data centre into a webform that the data paper author could edit/add to, and the resulting updated (or corrected) metadata could then be shared between the data centre and the journal. This would mean that the dataset author wouldn't have to provide information to the journal that had already been provided to the data centre, and would be able to improve/review the metadata held by the data centre at the same time as the data paper metadata is checked. (Note that this advanced linking is beyond the scope of the PREPARDE project, but is highlighted as potential future work.)
- The data journal has several places where it needs to communicate back to the data centre the results of reviews, when the data paper is published, citations of the data paper etc, so that further cross-links can be made to enrich the dataset. The data centre should link back to the data papers resulting from and papers that cite a particular dataset as well.

- The Dryad workflow as described at http://wiki.datadryad.org/Submission_System_Workflow is very much a software workflow, so doesn't describe the processes undergone or questions asked as part of the curator review (for example). It's also not clear where the Dryad workflow connects to the journal workflow, though we believe it comes in when a paper is submitted to a journal.
- Results of journal processes may impact the dataset as it's stored in a data centre, for example, results of data review might mean a correction has to be made and therefore a new version of a dataset (with a new DOI) has to be created and stored. It might be the case that a dataset would be withdrawn from the data centre, but the landing page for the DOI would be maintained. In this event, it would be appropriate for the data centre to notify the data journal that the data was no longer available, but it would be up to the journal to update the data paper and/or paper metadata to reflect this.
- A generalised workflow for dataset ingestion and publication is shown in figure 1.

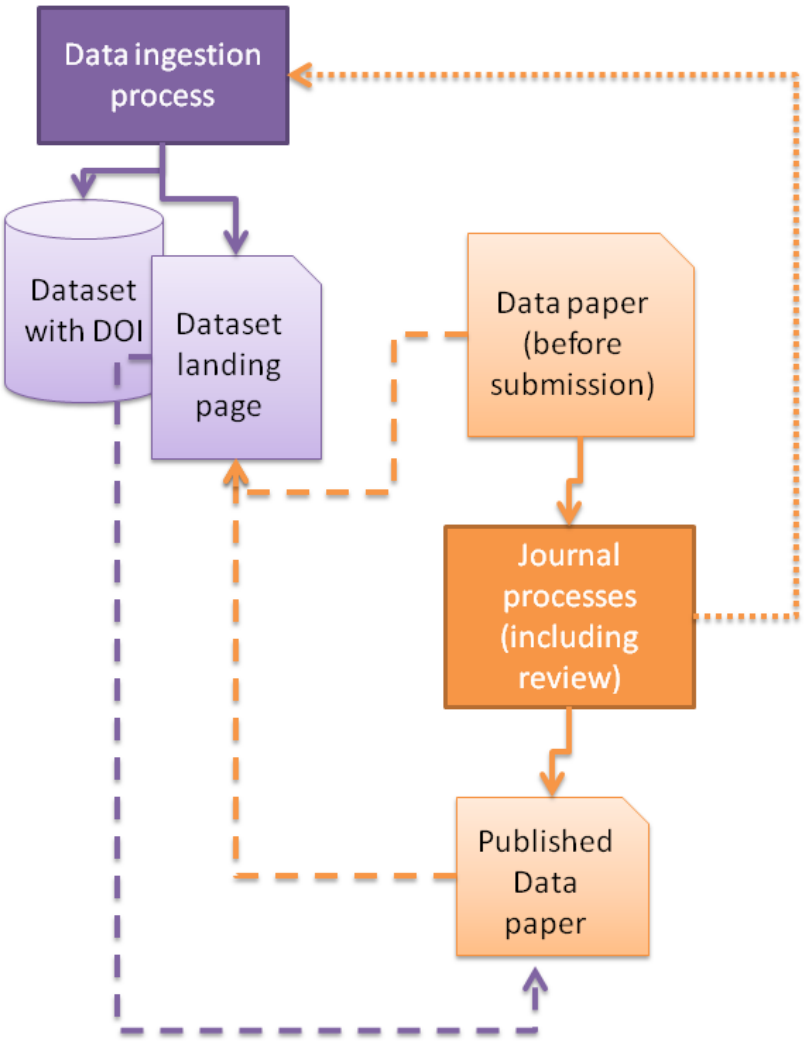
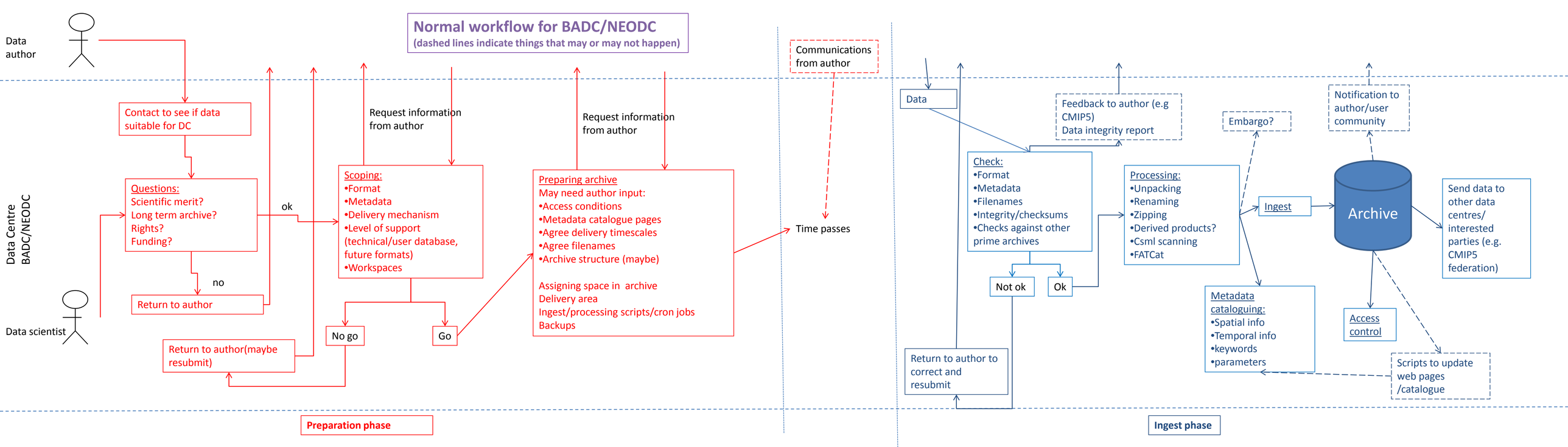
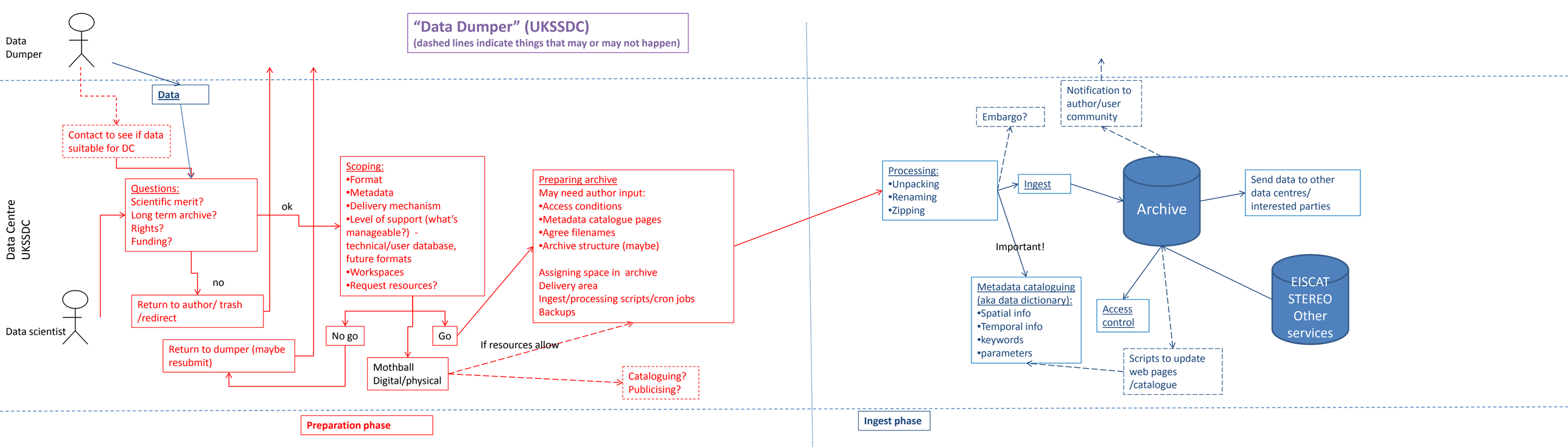
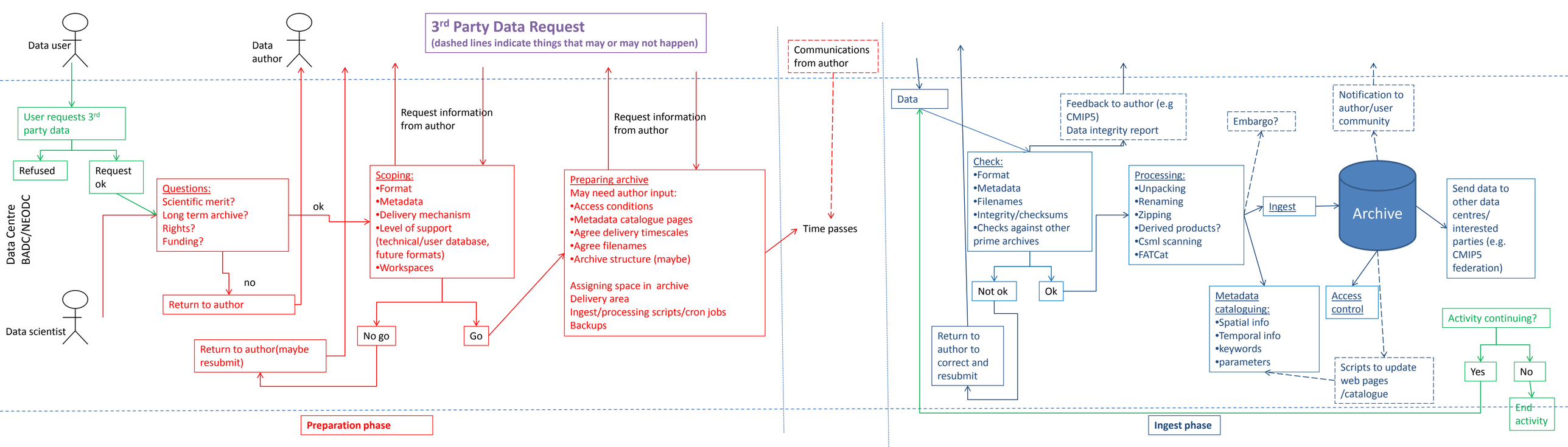


Figure 1: Generic data publication workflow. Dashed lines indicate linking (via URL) or citation (via DOI). Solid lines indicate the results or inputs into processes. Dotted line indicated where the results of a process need to be fed back into another process. Journal responsibilities are orange, data centre's are purple


Normal workflow for BADC/NEODC
(dashed lines indicate things that may or may not happen)
Data author
Data Centre BADC/NEODC
Data scientist
Contact to see if data suitable for DC
Questions:
Scientific merit?
Long term archive?
Rights?
Funding?
ok
no
Return to author
Request information from author
Scoping:
•Format
•Metadata
•Delivery mechanism
•Level of support (technical/user database, future formats)
•Workspaces
No go
Go
Return to author(maybe resubmit)
Request information from author
Preparing archive
May need author input:
•Access conditions
•Metadata catalogue pages
•Agree delivery timescales
•Agree filenames
•Archive structure (maybe)
Assigning space in archive
Delivery area
Ingest/processing scripts/cron jobs
Backups
Communications from author
Time passes
Preparation phase
Data
Feedback to author (e.g CMIP5)
Data integrity report
Check:
•Format
•Metadata
•Filenames
•Integrity/checksums
•Checks against other prime archives
Not ok
Ok
Return to author to correct and resubmit
Embargo?
Processing:
•Unpacking
•Renaming
•Zipping
•Derived products?
•Csml scanning
•FATCat
Ingest
Notification to author/user community
Archive
Send data to other data centres/ interested parties (e.g. CMIP5 federation)
Metadata cataloguing:
•Spatial info
•Temporal info
•keywords
•parameters
Access control
Scripts to update web pages /catalogue
Ingest phase


"Data Dumper" (UKSSDC)
(dashed lines indicate things that may or may not happen)
Data Dumper
Data
Contact to see if data suitable for DC
Data Centre UKSSDC
Questions:
Scientific merit?
Long term archive?
Rights?
Funding?
ok
no
Return to author/ trash /redirect
Data scientist
Return to dumper (maybe resubmit)
Scoping:
•Format
•Metadata
•Delivery mechanism
•Level of support (what's manageable?) - technical/user database, future formats
•Workspaces
•Request resources?
No go
Go
Mothball Digital/physical
Preparing archive
May need author input:
•Access conditions
•Metadata catalogue pages
•Agree filenames
•Archive structure (maybe)
Assigning space in archive
Delivery area
Ingest/processing scripts/cron jobs
Backups
If resources allow
Cataloguing?
Publicising?
Preparation phase
Embargo?
Notification to author/user community
Processing:
•Unpacking
•Renaming
•Zipping
Ingest
Archive
Send data to other data centres/ interested parties
Important!
Metadata cataloguing (aka data dictionary):
•Spatial info
•Temporal info
•keywords
•parameters
Access control
EISCAT
STEREO
Other services
Scripts to update web pages /catalogue
Ingest phase

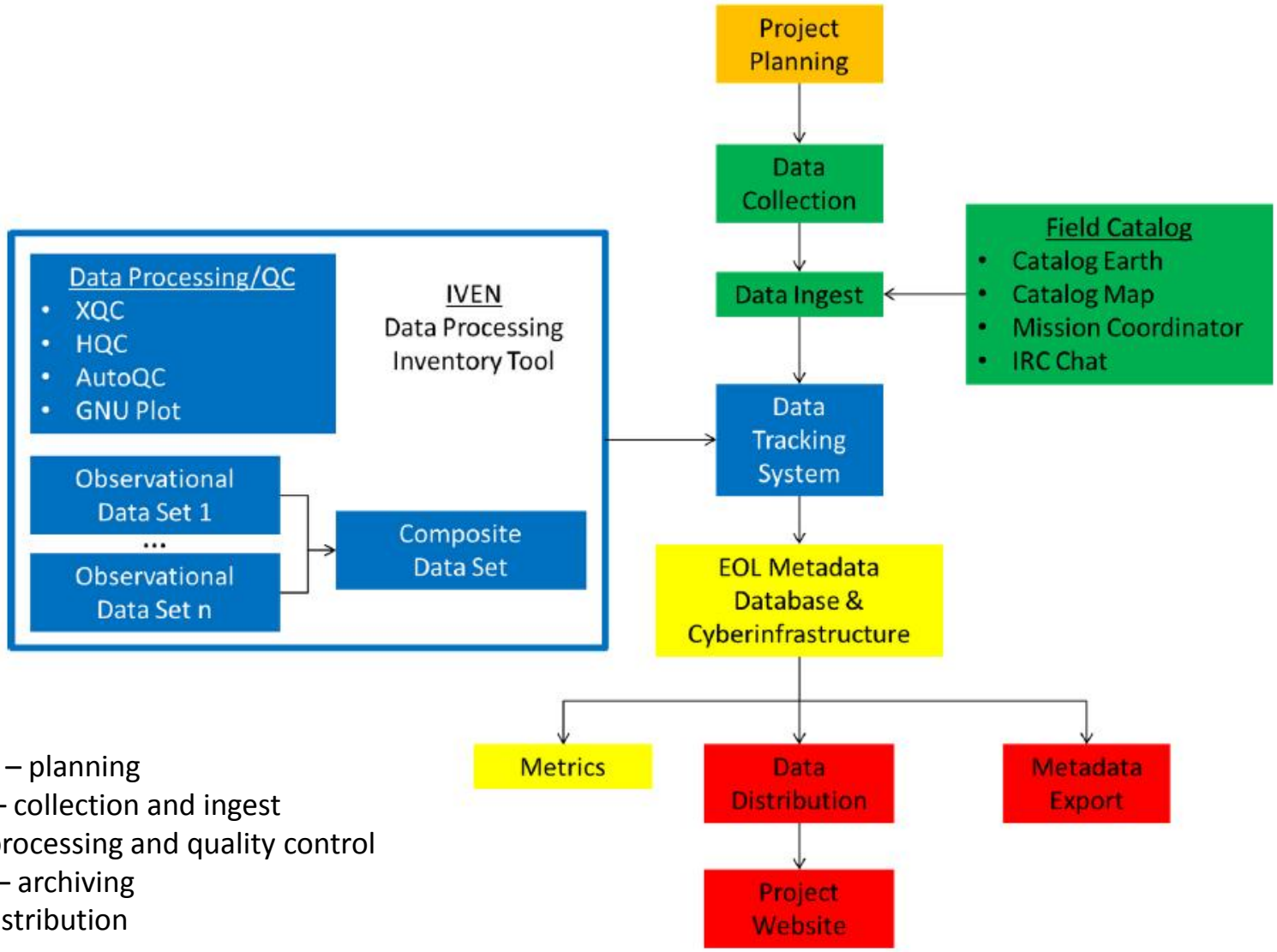

3rd Party Data Request
(dashed lines indicate things that may or may not happen)
Data user
Data author
Data scientist
Data Centre
BADC/NEODC
User requests 3rd party data
Refused
Request ok
Questions:
Scientific merit?
Long term archive?
Rights?
Funding?
ok
no
Return to author
Request information from author
Scoping:
•Format
•Metadata
•Delivery mechanism
•Level of support (technical/user database, future formats)
•Workspaces
No go
Go
Return to author(maybe resubmit)
Request information from author
Preparing archive
May need author input:
•Access conditions
•Metadata catalogue pages
•Agree delivery timescales
•Agree filenames
•Archive structure (maybe)
Assigning space in archive
Delivery area
Ingest/processing scripts/cron jobs
Backups
Communications from author
Time passes
Preparation phase
Data
Feedback to author (e.g CMIP5)
Data integrity report
Check:
•Format
•Metadata
•Filenames
•Integrity/checksums
•Checks against other prime archives
Not ok
Ok
Return to author to correct and resubmit
Embargo?
Notification to author/user community
Processing:
•Unpacking
•Renaming
•Zipping
•Derived products?
•Csml scanning
•FATCat
Ingest
Archive
Send data to other data centres/ interested parties (e.g. CMIP5 federation)
Metadata cataloguing:
•Spatial info
•Temporal info
•keywords
•parameters
Access control
Scripts to update web pages /catalogue
Activity continuing?
Yes
No
End activity
Ingest phase

NCAR Earth Observing Laboratory (EOL): Computing, Data, and Software Facility

Workflow Diagrams
Oct. 2012

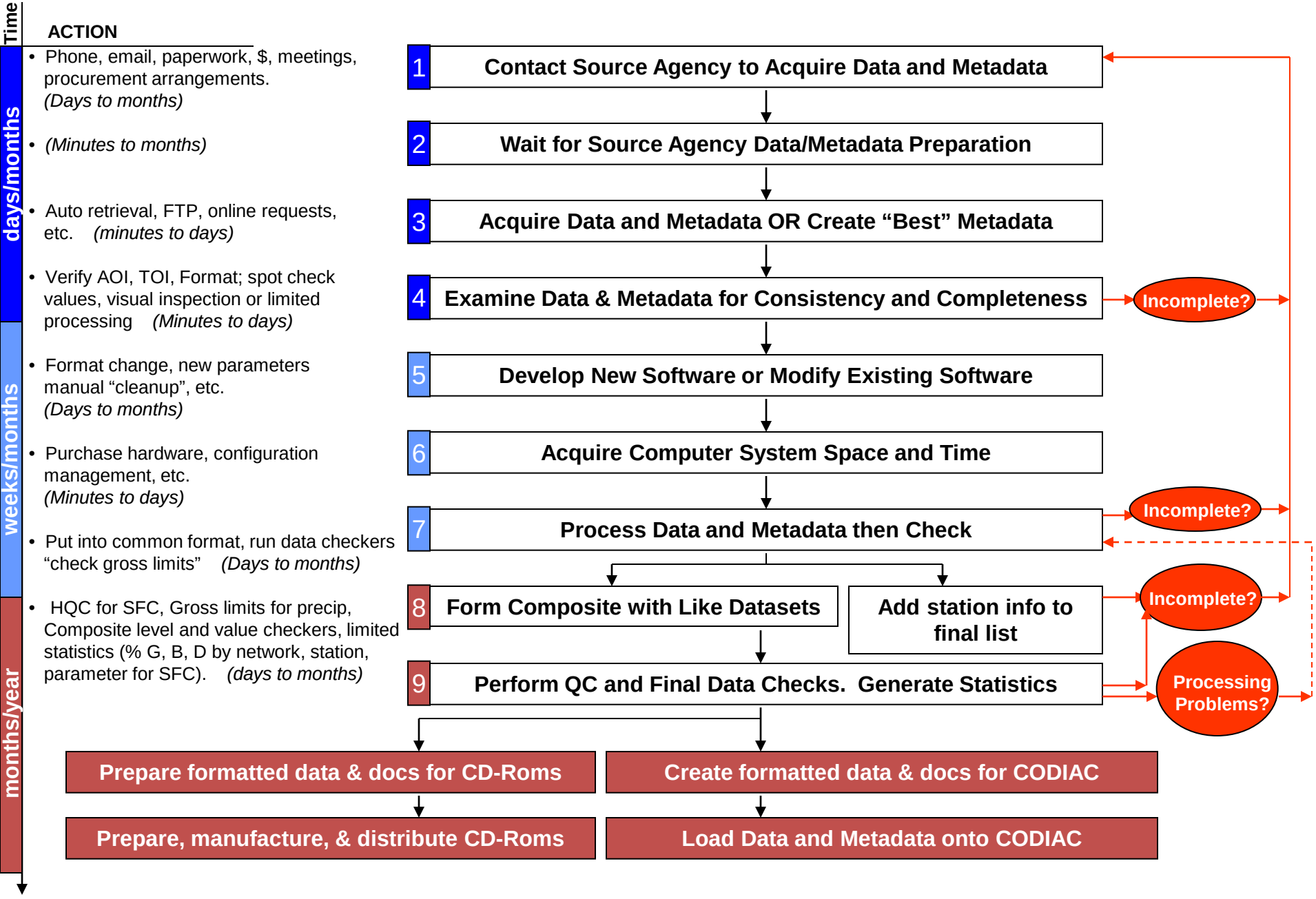
Compiled by
Matt Mayernik, Steve Williams, & Mike Daniels (NCAR)
for the PREPARDE project

1. NCAR EOL Data Management Group Workflow

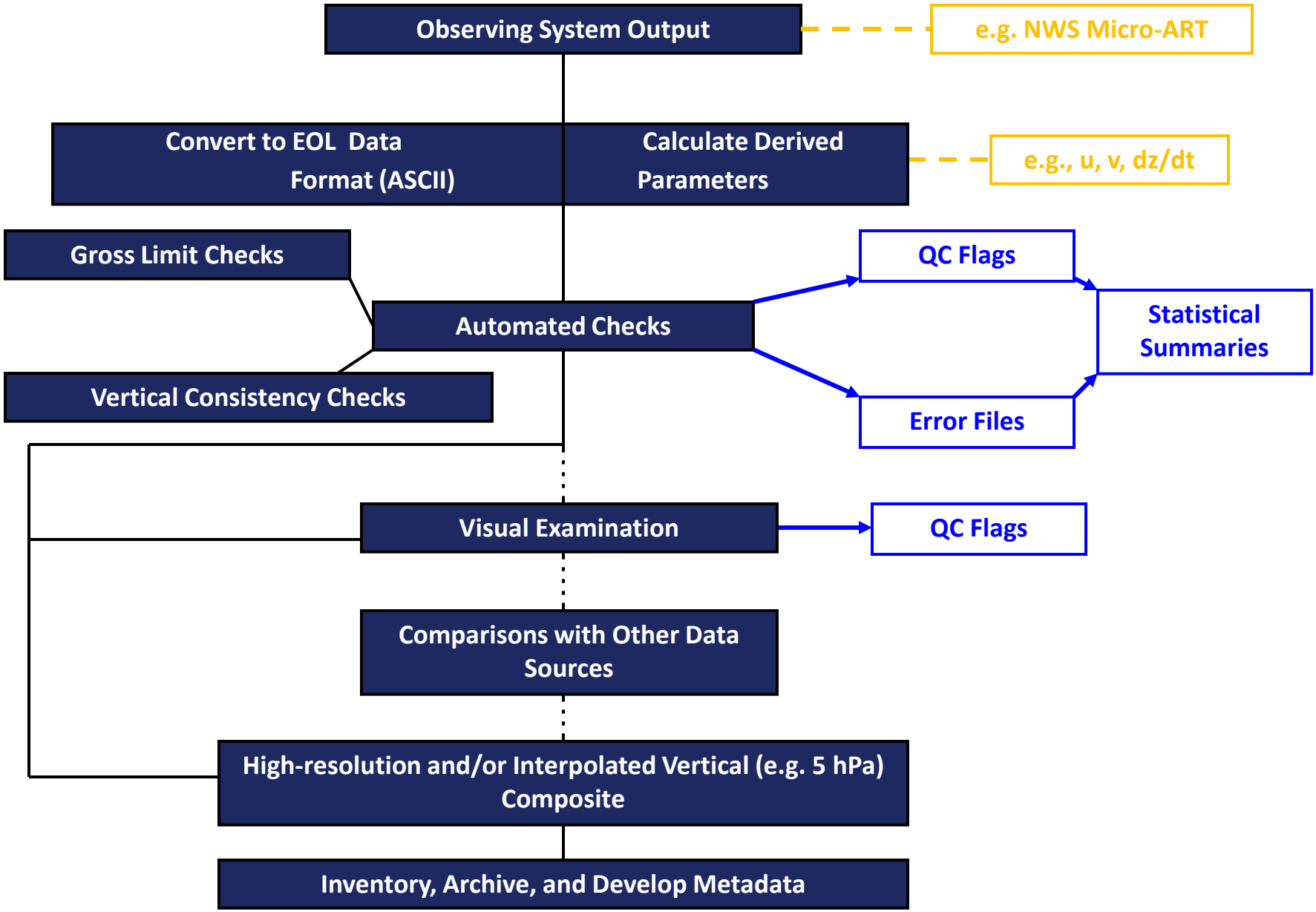


Colors:
- Orange – planning
- Green – collection and ingest
- Blue - processing and quality control
- Yellow – archiving
- Red - distribution

2. GENERAL DATASET PROCESSING LIFE CYCLE



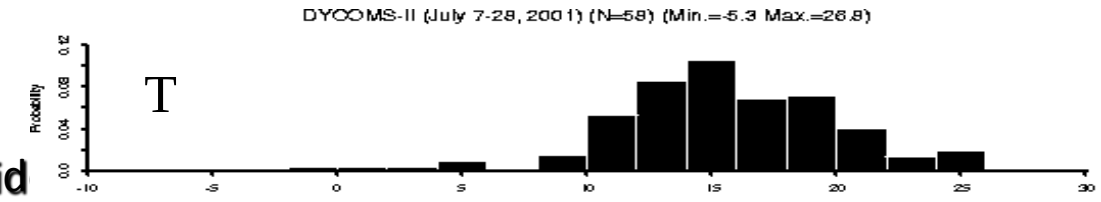
3. NCAR/EOL Atmospheric Sounding Processing Procedures

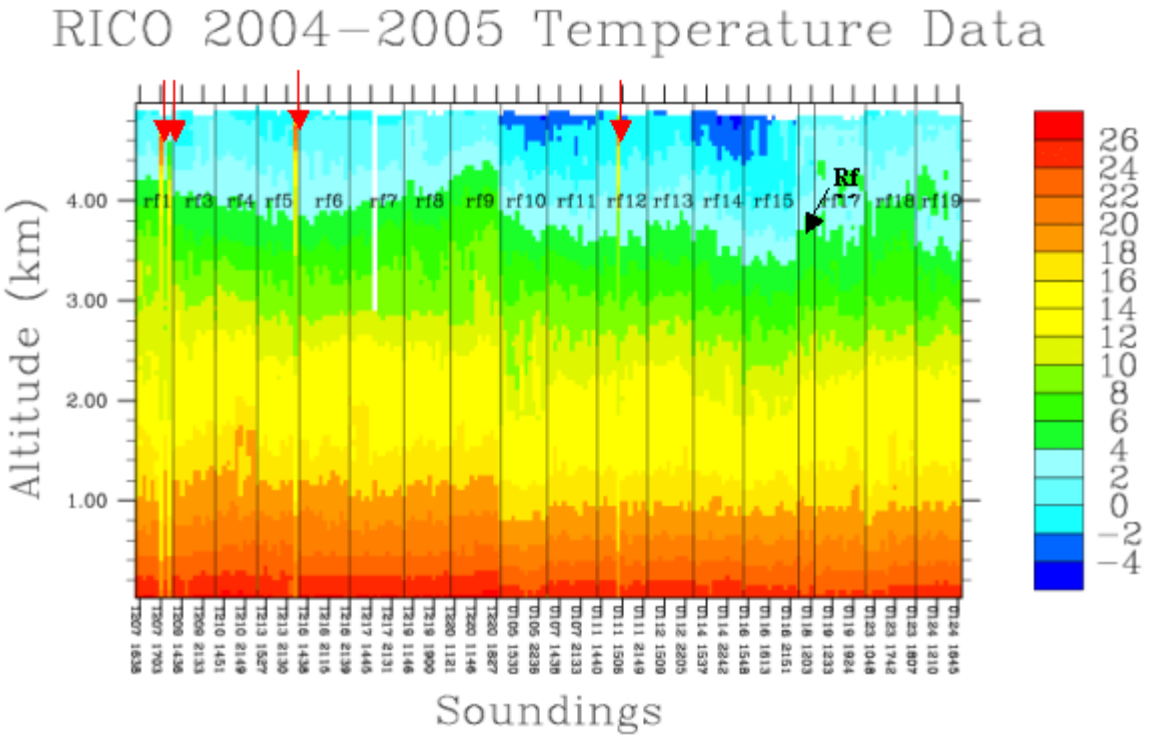


4. EOL Quality Control of Dropsonde Data

1. In flight data inspection
2. ASPEN
3. Individual Skew-t Examination
4. Histograms of PTU and Wind
5. Time series of PTU and Wind
6. Comparisons with other data

Provid





Acronyms

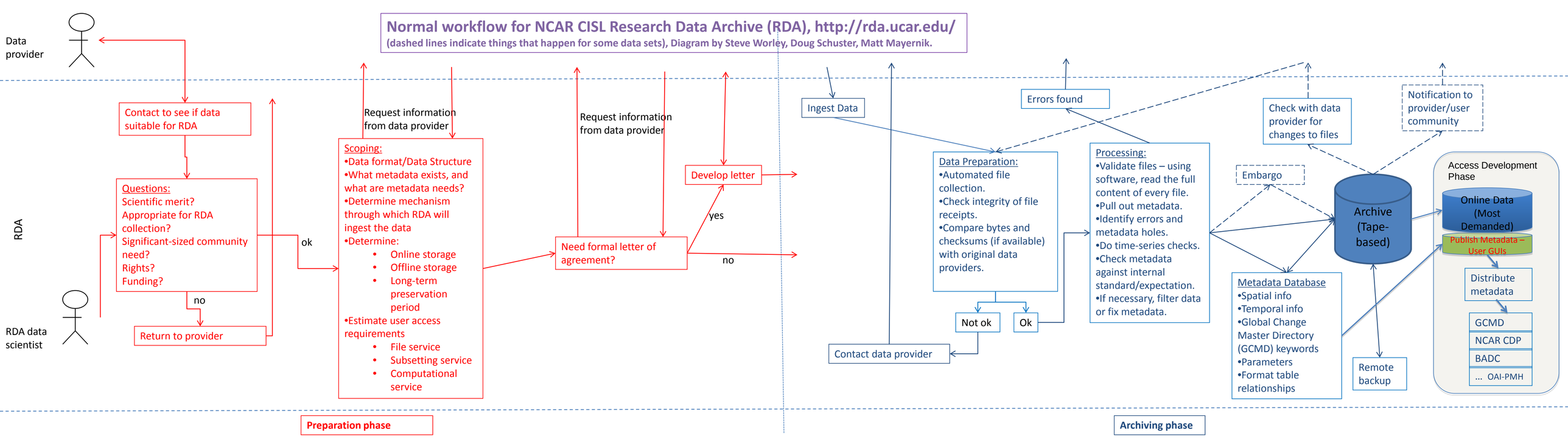
- AOI - Area of Influence (areal subset)
- ASPEN - Atmospheric Sounding Processing Environment
- AutoQC – EOL “Automated Quality Control“ tool
- CODIAC – EOL data archive access tool
- G, B, D – Flags used in the data: G=Good, B=Bad, D=Dubious
- HQC - Horizontal Quality Control
- IRC Chat - Internet Relay Chat
- IVEN – EOL Inventory Tool used as part of EOL’s Data Tracking System
- NWS Micro-ART - National Weather Service Microcomputer Automatic Radio-theodolite
- PTU - pressure, temperature, and humidity
- SFC - Surface Meteorological and Radiation Data Set
- TOI - Time of Influence (spatial subset)
- XQC – EOL “Xwindows Quality Control” tool

Diagram sources

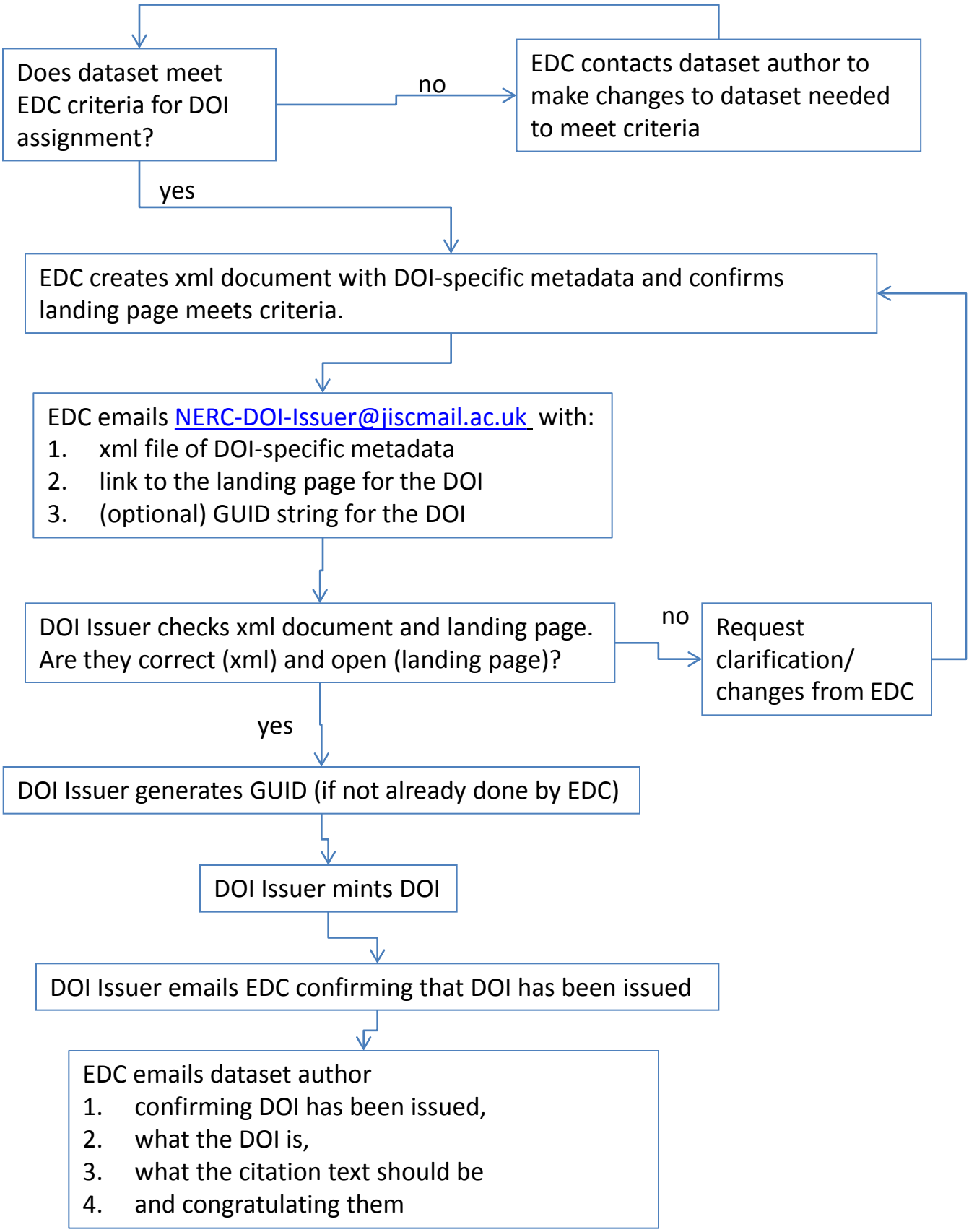
- Diagram 1 from:
  - Christopher Eaker. *Data Audit and Analysis: Mapping the Data Workflow from Ingest to Archive*. Unpublished internal document, prepared for NCAR Earth Observing Lab, July 20, 2012.
- Diagram 2 adapted from:
  - Steven F. Williams, Scot M. Loehrer, Linda E. Cully, Darren R. Gallant, Janine Goldstein, and Don Stott. *IHOP-2002 Data Archive and Development of Composite Data Sets.* IHOP-2002 Spring Science Workshop, Boulder, CO, March 2003. www.eol.ucar.edu/dir_off/projects/2002/IHOPwsMar03/loehrer.ppt
- Diagram 3 and 4 adapted from:
  - Steve Williams, Chris Webster, and Dennis Flanigan. *Data Formats at EOL*. Joint EOL/Unidata Seminar. Boulder, CO, May 29, 2007. www.unidata.ucar.edu/Presentations/UPCsemseries/EOL-Unidata_Formats_0507.ppt

Normal workflow for NCAR CISL Research Data Archive (RDA), http://rda.ucar.edu/

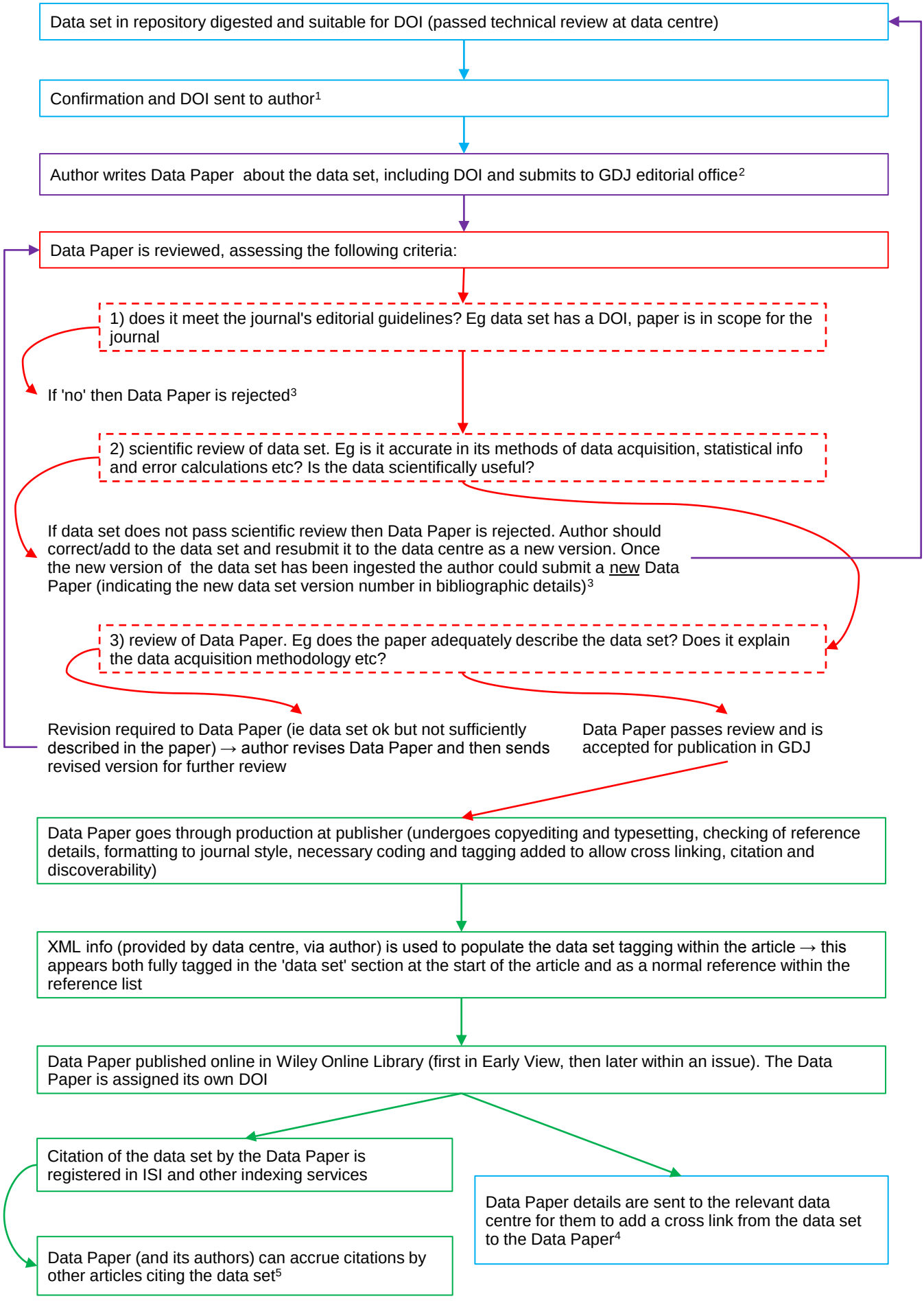
(dashed lines indicate things that happen for some data sets), Diagram by Steve Worley, Doug Schuster, Matt Mayernik.



Flowchart for minting DOIs for NERC environmental data centres (EDC) – December 2011



Geoscience Data Journal
Data Paper workflow



GDJ Workflow notes

[1] Wiley need to specify format/content of this info, eg an XML file with necessary bibliographic info would be ideal.

[2] We should require the author to send the XML file generated by the data centre along with the submission, to extract the necessary bibliographic details for the data set from that.

[3] Need to communicate rejection to Data Centre so they flag as 'rejected'.

[4] An API needs to be developed for this to allow easy integration of Data Paper details into data centre on publication.

[5] Possible future development would be to pass these citations on to the data centre so that they can also be shown as cross links from the original data set in the data centre.

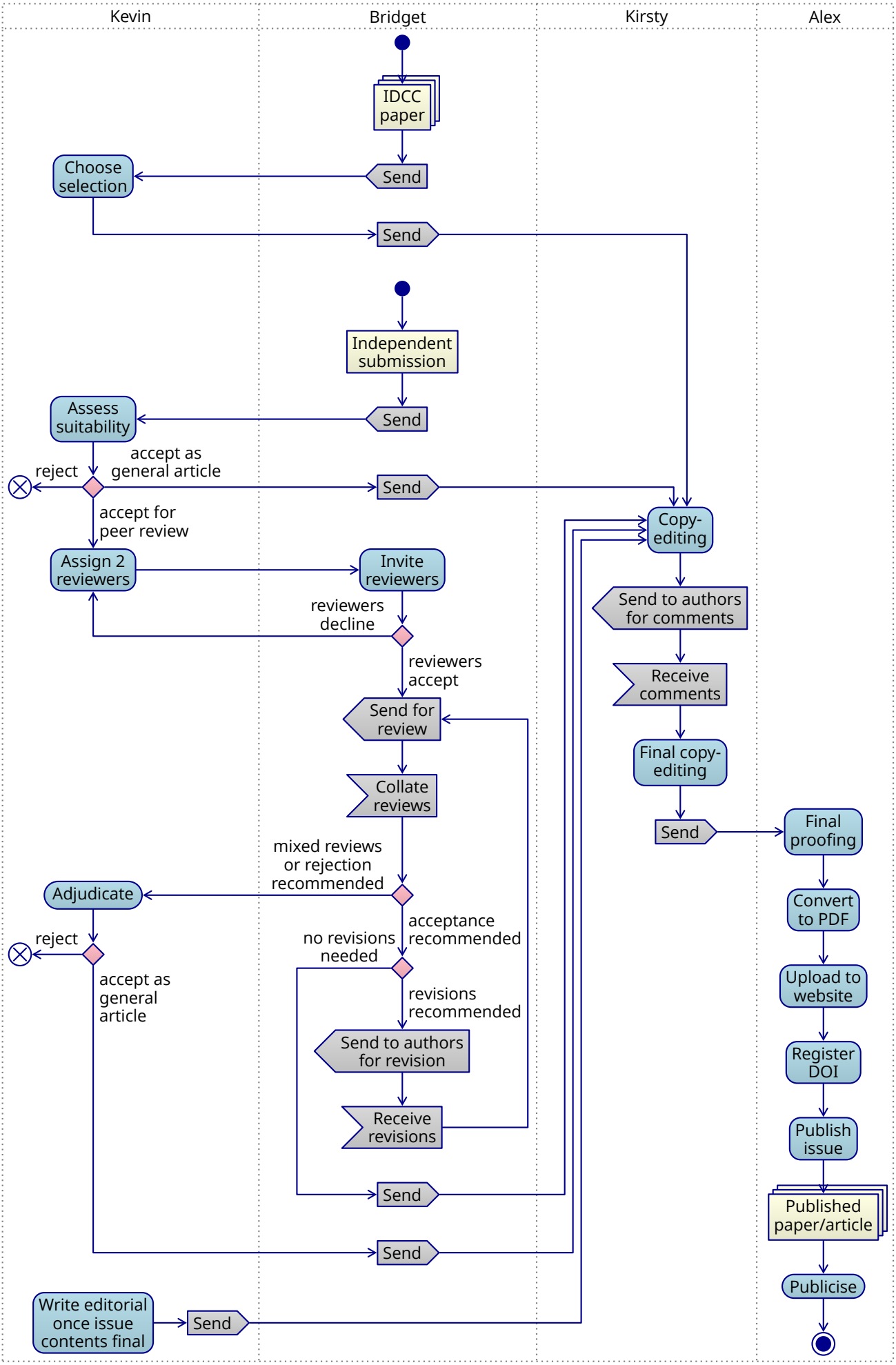
Key to workflow diagram:

Blue: stages performed by data centre

Purple: stages performed by Data Paper author

Red: stages performed by GDJ editorial office

Green: stages performed by publisher


Kevin
Bridget
Kirsty
Alex
IDCC paper
Send
Choose selection
Send
Independent submission
Send
Assess suitability
reject
accept as general article
Send
accept for peer review
Assign 2 reviewers
Invite reviewers
reviewers decline
reviewers accept
Send for review
Collate reviews
mixed reviews or rejection recommended
Adjudicate
reject
accept as general article
no revisions needed
acceptance recommended
revisions recommended
Send to authors for revision
Receive revisions
Send
Send
Copy-editing
Send to authors for comments
Receive comments
Final copy-editing
Send
Final proofing
Convert to PDF
Upload to website
Register DOI
Publish issue
Published paper/article
Publicise
Write editorial once issue contents final
Send