

# La reproductibilité en science et la RDA: Zoom sur le groupe de travail Reproducibility Checklist

*Christian Pagé, CERFACS/CECI/IRD*

## Contexte (1/3)

- Groupe de Travail (WG) RDA "Reproducibility Checklist" : "Recognised and Endorsed" !
- **Chairs** : Claire Austin, Jiban K. Pal, **Christian Pagé**, Leyla Jael Castro, Daniela Gawehtns
- **Objectif** : Liste de contrôle standardisée et indépendante de toute discipline pour l'évaluation de la reproductibilité

L'objectif d'une liste de contrôle standardisée et indépendante de toute discipline pour l'évaluation de la reproductibilité est de promouvoir l'intégrité scientifique en fournissant un cadre pratique aux chercheurs, scientifiques et organisations de toutes disciplines afin d'évaluer et de documenter de manière systématique la reproductibilité numérique dans les sciences et la recherche quantitatives et qualitatives.

# Le Constat — L'Iceberg de la Reproductibilité (2/3)

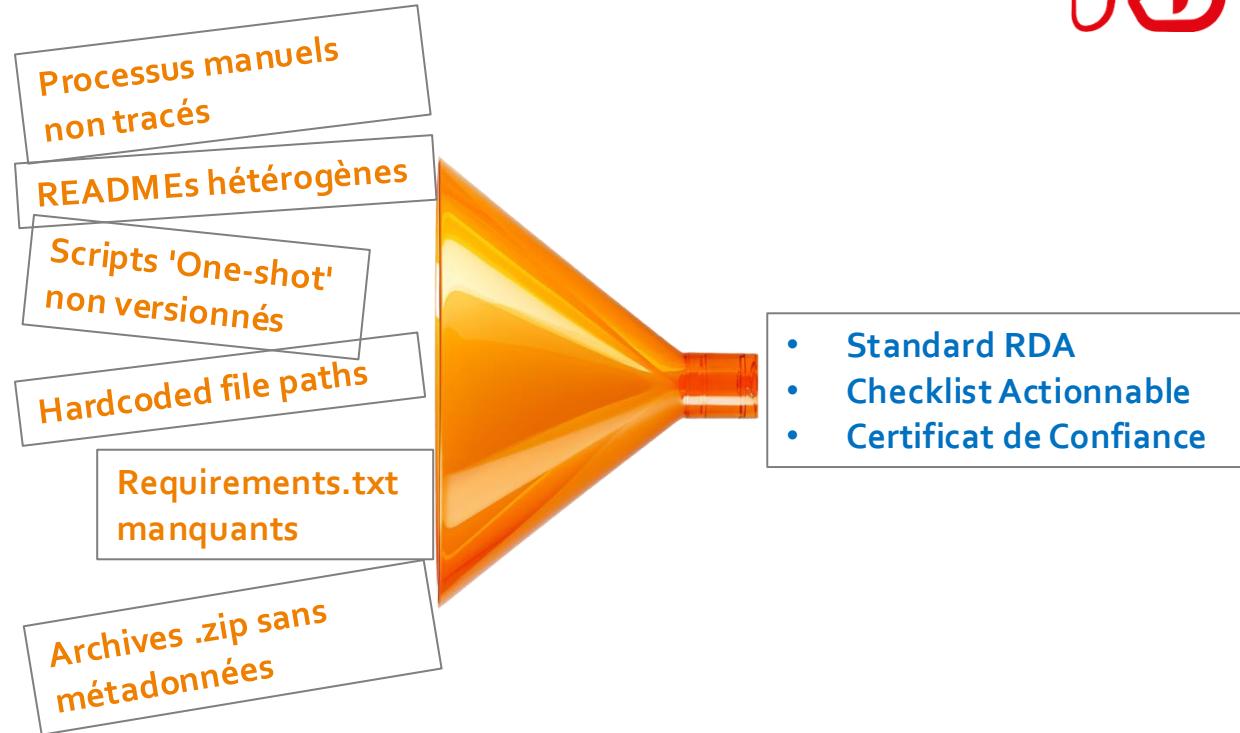
- Le principe de données ouvertes n'est pas une fin en soi.
- **Le défi** : "Six mois plus tard, mon code ne tourne plus".
- **Besoin** : Passer d'un dépôt passif à un objet de recherche "vivant".

## Open Data



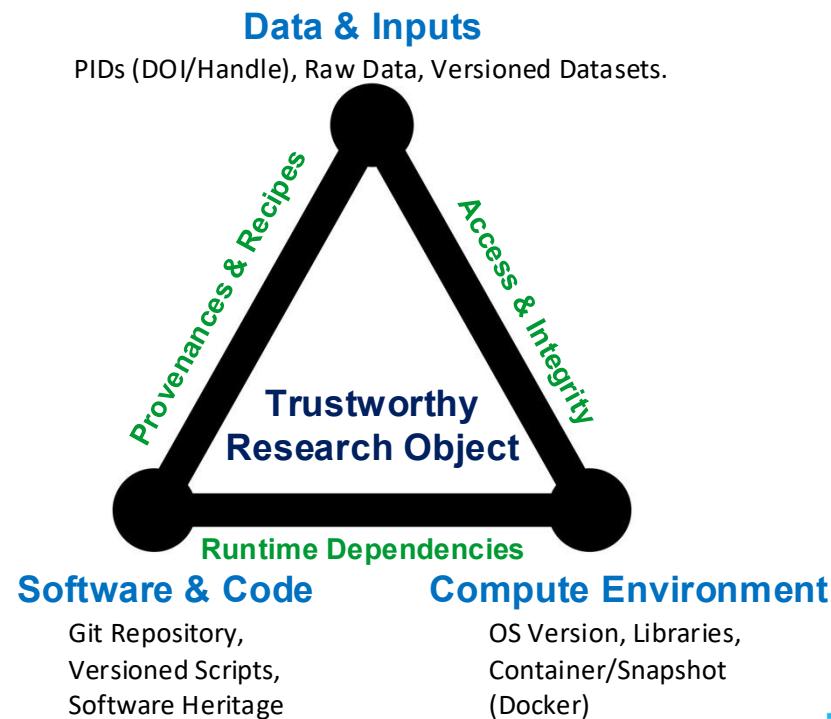
# La mission du WG RDA Reproducibility Checklist (3/3)

- **Standardiser** les critères de reproductibilité
- Dépasser les principes théoriques (FAIR) pour l'**implémentation**
- Développer un protocole de *vérification* pour les infrastructures.



# SoW Axe 1 — Le Trépied de la Reproductibilité (1/4)

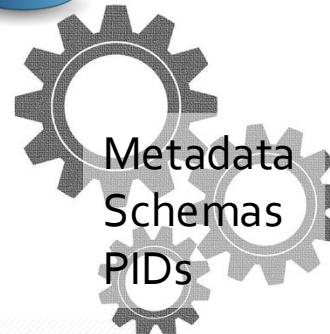
- **Données** : Inputs identifiés (PIDs)
- **Logiciel** : Code source versionné et archivé
- **Environnement** : Paramètres système et bibliothèques (Runtime)
- **Note** : *L'absence d'un pilier rend la preuve scientifique caduque*



# La liaison technique (Métadonnées) (2/4)

- **Liaison** : Utilisation de Schema.org, CodeMeta
- **Traçabilité** : Identifier quel code a produit quel résultat, Provenance (PROV-O)
- **PIDs** : DOIs pour les données et Software Heritage IDs pour le code

Données



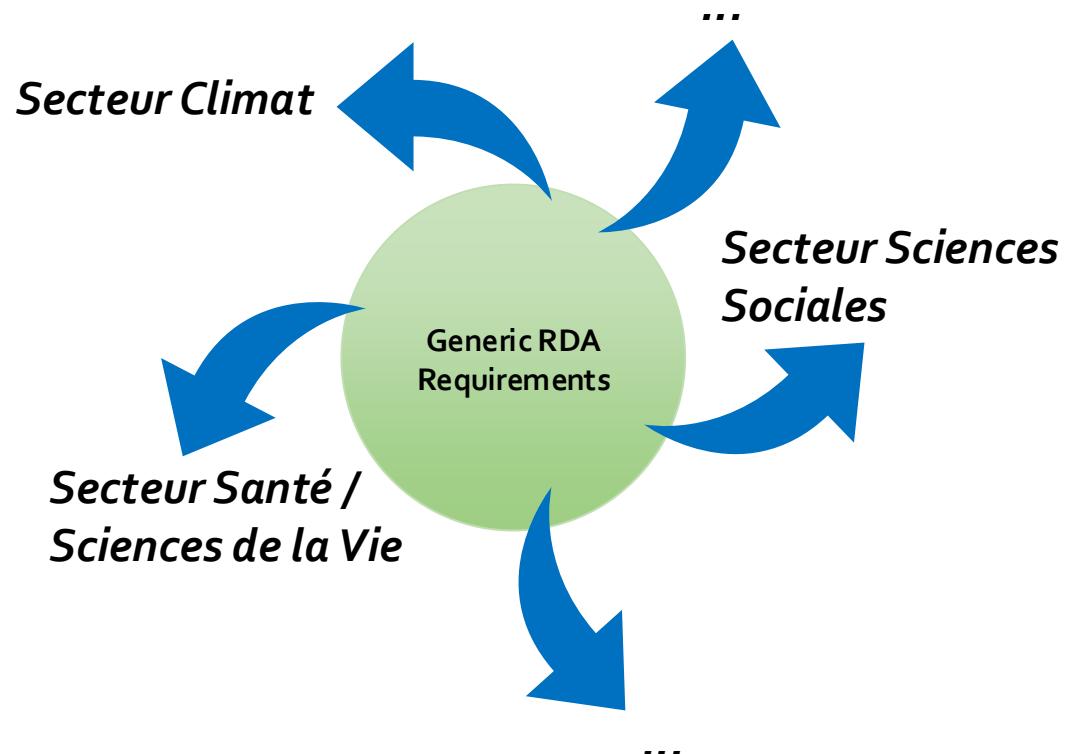
Environnement



Codes

## SoW Axe 2 – Les Crosswalks (L'interopérabilité) (3/4)

- **Objectifs** : Ne pas réinventer la roue, mais bâtir des ponts
- **Le principe des Crosswalks** : Mapper les exigences de la Checklist sur les métadonnées déjà utilisées par les communautés
- **Bénéfices** : Permettre à une infrastructure de recherche d'un domaine de lire des données d'un autre domaine



# Sow Axe 2 – Les Crosswalks (L'interopérabilité) (4/4)

Time consuming

edit metadata

A metadata form

An example of basic AMI metadata entry.

Create Date\*

2018-09-13T08:06:02.351-04:00

Name

The MediaPreserve

URL

<https://mirlyn.lib.umich.edu/Record/014616558/>

Format

WAVE

Filename

auam-3915091568454-001.wav

Download

Open Sandbox

Error-prone

experimental data

time  $t$  (min) conc.  $\rho$  ( $\mu\text{g}/\text{mL}$ )

time $t$ (min)	conc. $\rho$ ( $\mu\text{g}/\text{mL}$ )
20	70.4
40	43.3
60	30.8
80	22.6
100	16.8
120	12
140	8.92
160	6.68
180	4.85
200	3.91
220	2.72
240	1.92

Average: 74.37063333 Count: 24 Sum: 1794.8 100%

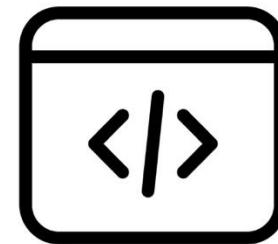
Ready



- **Zéro paperasse** : la checklist doit être automatisable
- **Vérification** par les infrastructures de calcul (HPC/Cloud)
- **Niveaux de conformité** : Bronze, Silver, Gold (selon la rigueur)



Machine-Actionable



Scalable

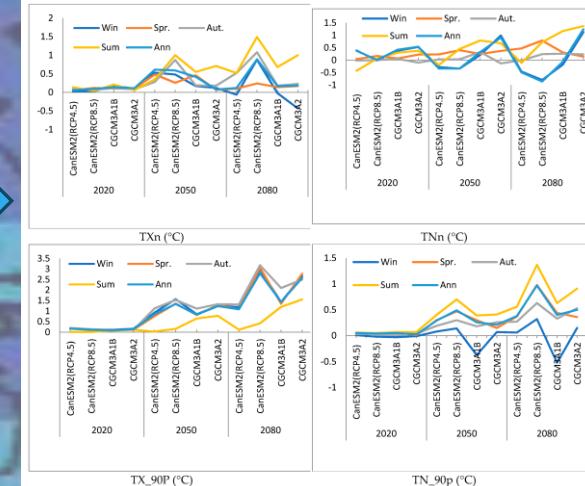
Infrastructure-ready

# Exemple d'application: La recherche en climat (1/4)

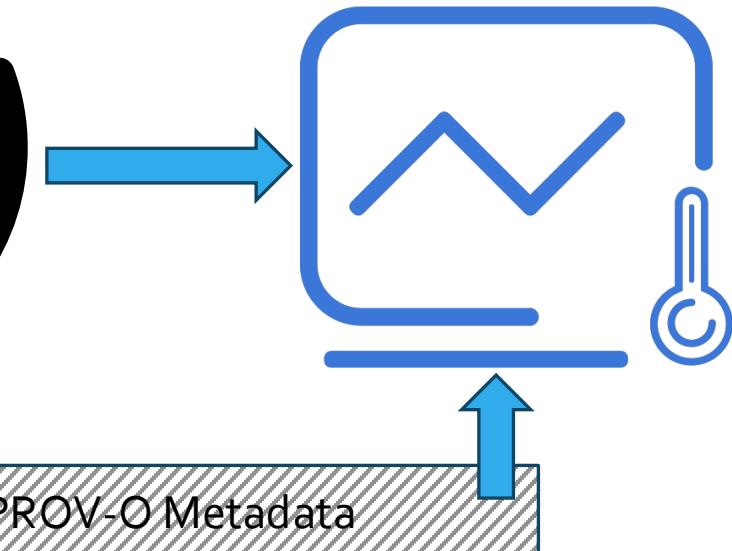
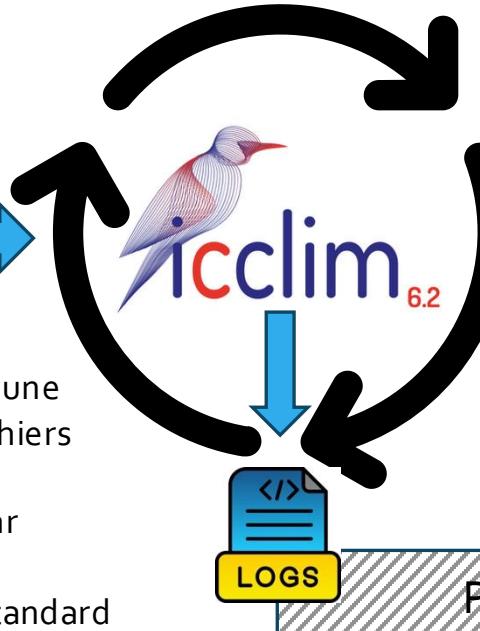
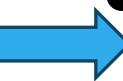
The Big Data  
Challenge  
Petabytes

Multi-Step Workflows

- **Échelle massive:** Des volumes de données mondiaux (CMIP, CORDEX) à gérer et à tracer.
- **Workflows multi-étapes:** De la donnée brute aux indicateurs d'impact
- **Responsabilité sociétale:** Des données qui servent aux politiques d'adaptation (besoin de confiance).
- Cas d'usage lourd techniquement



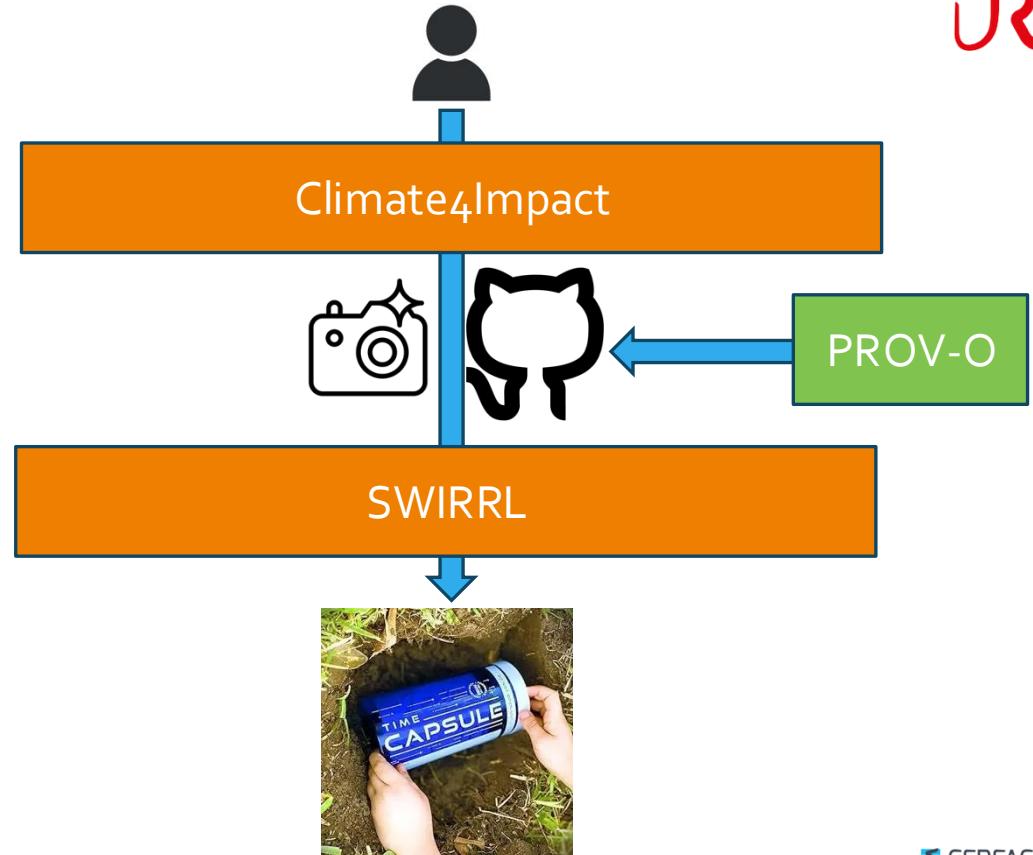
## Exemple d'application: icclim & Provenance (2/4)



- **État actuel** : Capture automatique d'une provenance dans les attributs des fichiers
- **Limites** : Documentation encore fragmentée et difficilement lisible par d'autres systèmes.
- **Projet technique** : Implémenter le standard PROV-O pour structurer cette information.
- **Objectif** : Rendre la provenance de l'indicateur universelle et exploitable par des outils d'audit automatiques.

# Technologie – Climate4Impact (C4I) (3/4)

- **Plateforme C4I** : Orchestration des calculs à distance sur les données de l'ESGF
- **Technologie SWIRRL** : Gestion dynamique des environnements de travail (JupyterLab, outils). Enregistrement automatique du workflow complet au standard PROV-O.
- **Snapshots reproductibles** : Capture instantanée de l'état complet du workflow
- **Pérennité** : Le même environnement peut être relancé à l'identique dans le futur



# Technologie – Climate4Impact (C4I) (4/4)

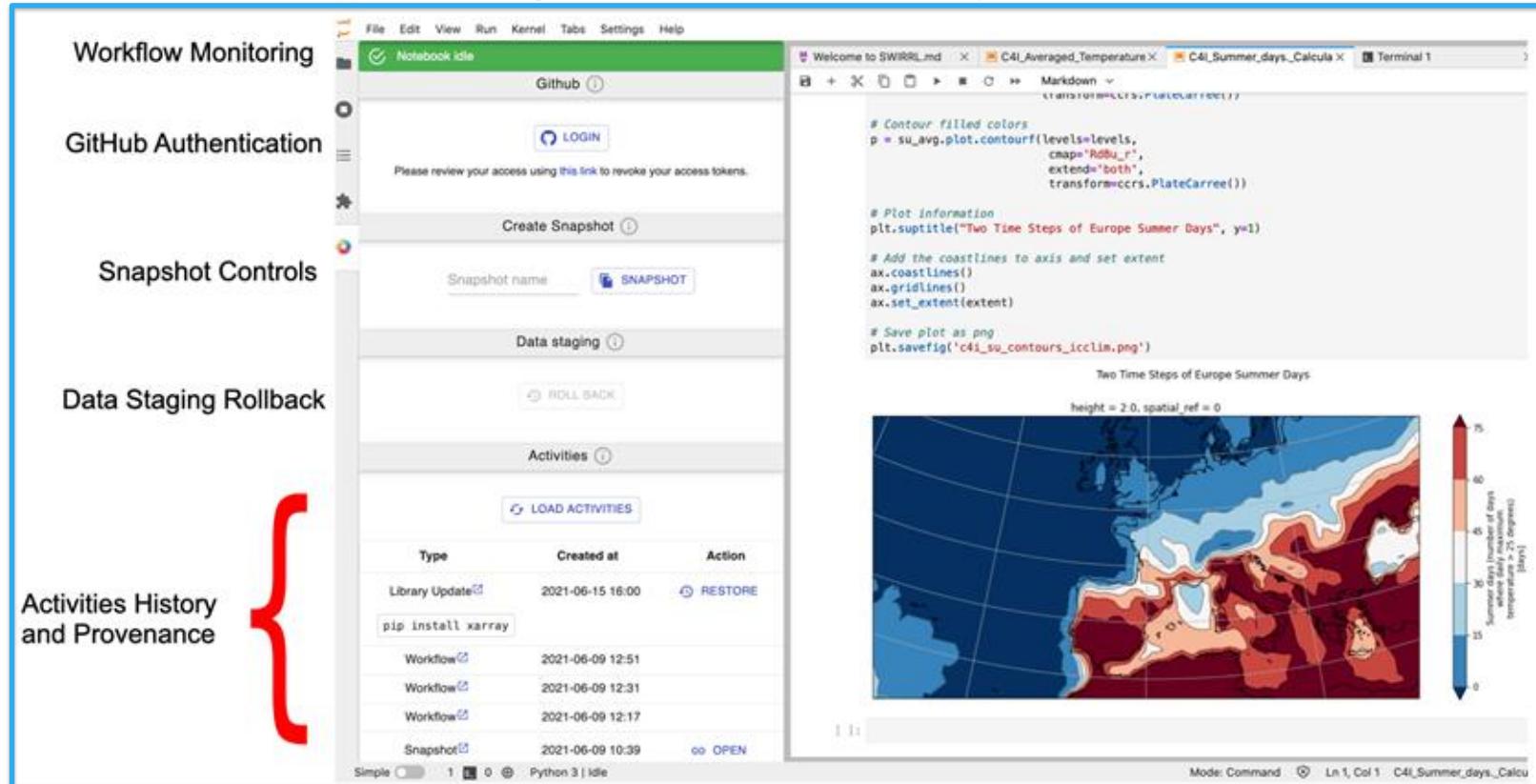
Workflow Monitoring

GitHub Authentication

Snapshot Controls

Data Staging Rollback

Activities History and Provenance



The screenshot displays the Climate4Impact (C4I) interface, which integrates several tools for scientific workflow management and data analysis. On the left, a vertical sidebar lists five main sections: 'Workflow Monitoring', 'GitHub Authentication', 'Snapshot Controls', 'Data Staging Rollback', and 'Activities History and Provenance'. A large red brace groups the 'Activities History and Provenance' section and the 'Data Staging Rollback' section. The main workspace is divided into three main areas: 1) A 'Workflow Monitoring' dashboard showing a 'Notebook idle' status with a GitHub authentication step (a 'LOGIN' button and a link to review access tokens). 2) A 'Data Staging Rollback' section with a 'Create Snapshot' button, a 'Snapshot name' input field, and a 'ROLL BACK' button. 3) A 'Activities' section with a 'LOAD ACTIVITIES' button and a table listing activities: 'Library Update' (Created at 2021-06-15 16:00, Action: RESTORE), 'pip install xarray' (Created at 2021-06-09 12:51), 'Workflow' (Created at 2021-06-09 12:31), 'Workflow' (Created at 2021-06-09 12:17), and 'Snapshot' (Created at 2021-06-09 10:39, Action: OPEN). The right side of the interface shows a Jupyter Notebook environment with a terminal tab. The notebook code generates a map titled 'Two Time Steps of Europe Summer Days' showing the number of days with temperatures above 25 degrees Celsius. The map uses a color scale from blue (0 days) to red (75 days). The code includes: 

```
# Contour filled colors
p = su_avg.plot.contourf(levels,
                           cmap='RdBu_r',
                           extend='both',
                           transform=ccrs.PlateCarree())

# Plot information
plt.suptitle("Two Time Steps of Europe Summer Days", y=1)

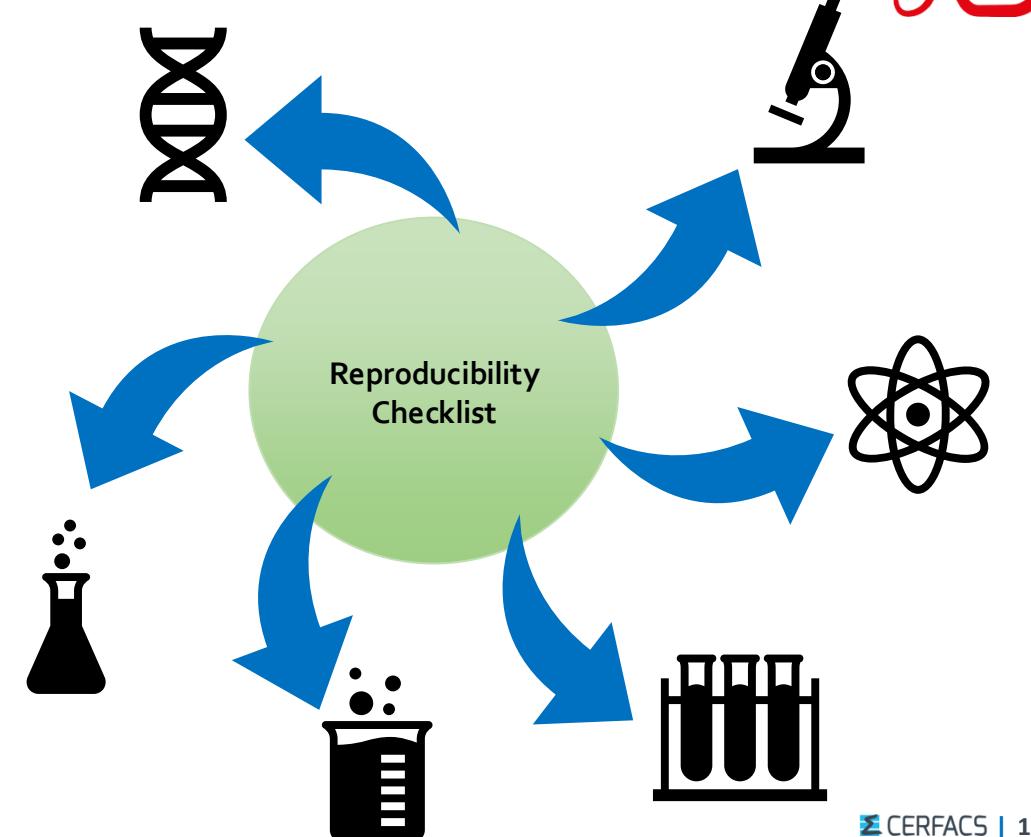
# Add the coastlines to axis and set extent
ax.coastlines()
ax.gridlines()
ax.set_extent(extent)

# Save plot as png
plt.savefig('c4i_su_contours_icclim.png')
```

 The map shows high concentrations of summer days in Southern Europe and parts of Africa, with a color bar indicating values from 0 to 75 days.

# La Reproducibility Checklist: Un cadre agnostique au domaine (1/3)

- **Indépendance disciplinaire** : La reproductibilité est une exigence méthodologique, pas une thématique
- **Dénominateur commun** : Quelle que soit la discipline, les questions restent les mêmes
  - Qui ? (Acteurs et provenance)
  - Quoi ? (Données et versions)
  - Comment ? (Code et environnement)
- **Interopérabilité**: Faciliter la réutilisation des résultats par des tiers sans connaissance préalable de la "cuisine" interne du domaine



# Un chantier de 18 mois (2/3)

- **Le départ d'un cycle** : Le groupe de travail (WG) RDA vient de lancer ses travaux pour une durée de 18 mois.
- **Rien n'est figé** : C'est le moment idéal pour influencer le standard avant qu'il ne soit finalisé.
- **Besoin de testeurs** : Nous aurons besoins de workflows variés pour tester la robustesse de la Checklist.
- **Objectif** : Aboutir à une checklist et un outil qui soit le reflet des besoins réels des laboratoires.

## Timeline 18 mois

**Étape 1**  
Définition des critères et premiers tests.

**Étape 2**  
Confrontation aux cas d'usage

**Étape 3**  
Finalisation du standard RDA

# Conclusion : Vers une culture de la reproductibilité (3/3)

- **Le Trépied Indissociable**

- Pas de reproductibilité sans le lien : Données + Code + Environnement.
- L'environnement (OS, bibliothèques) est le pilier le plus fragile.

- **L'Automatisation est une réalité**

- La technologie (PROV-O, ...) permet de capturer les informations sans alourdir le travail du chercheur.
- L'infrastructure et les outils doivent soutenir cette automatisation.

- **Un Standard en construction (18 mois)**

- La Reproducibility Checklist : un cadre agnostique pour sortir des silos disciplinaires.
- C'est le moment pour tous de contribuer à construire ce futur standard.



**Garantir la reproductibilité, c'est un pilier essentiel de la recherche scientifique**