

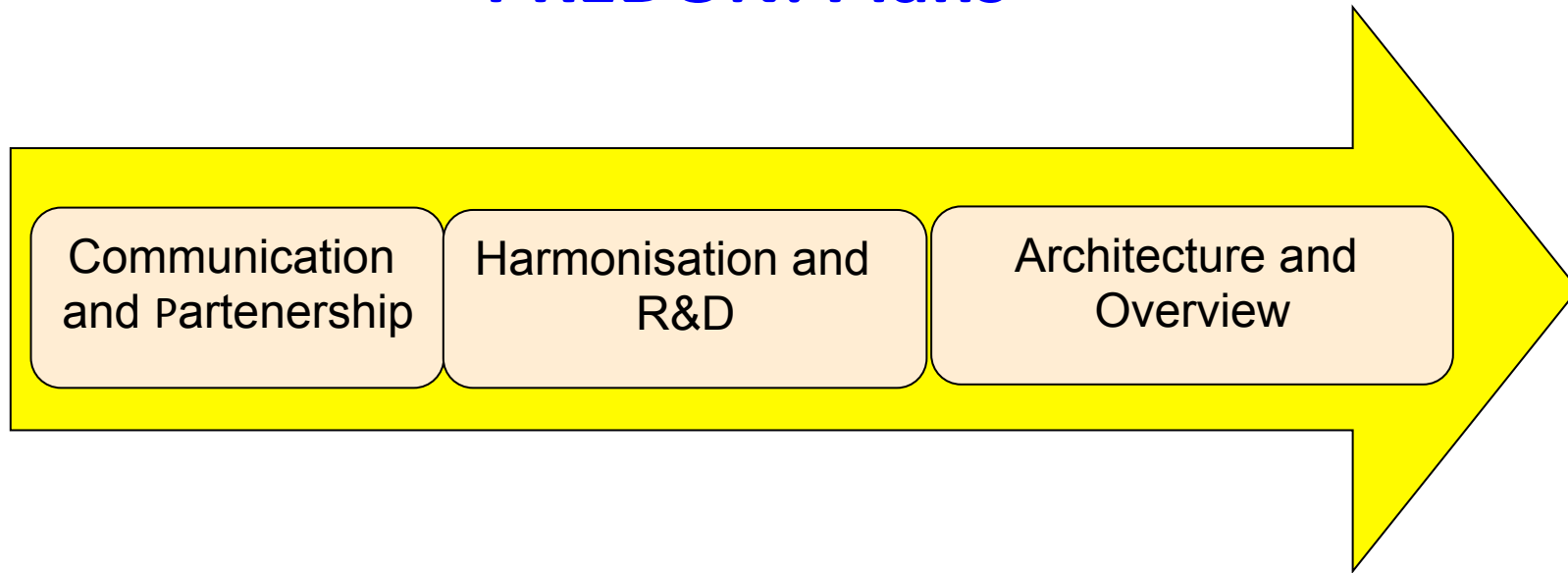
PREDON

A project on Scientific Data Preservation in France
within MASTODONS multi-disciplinary program

MASTODONS

- Multi-disciplinary Department of CNRS launched a call in April 2012:
 - Data exceeds storage
 - More science in Data
- Possible Directions:
 - Stockage et gestion de données (par exemple, dans le Cloud), sécurité, confidentialité.
 - Calcul intensif sur des grands volumes de données, parallélisme dirigé par les données.
 - Visualisation de grandes masses de données.
 - Extraction de connaissances, datamining et apprentissage.
 - Qualité des données, confidentialité et sécurité des données.
 - Problèmes de propriété, de droit d'usage, droit à l'oubli.
 - **Préservation/archivage des données pour les générations futures.**
 - **PREDON (PREservation des DONnees)**
- MASTODONS is likely to be evolved in a national program around big data

PREDON: Plans



- Short term (2013/2014): **Communication and partenership**
 - Enlarge the community
- Medium term (2014/2015) : **Harmonisation and R&D**
 - Communication: exchanges and workshops
 - Demonstrator acces and préservation
- Long term (2016) **Architecture and overview**
 - “Observatoire National des Données Scientifiques”

PREDON: Challenges

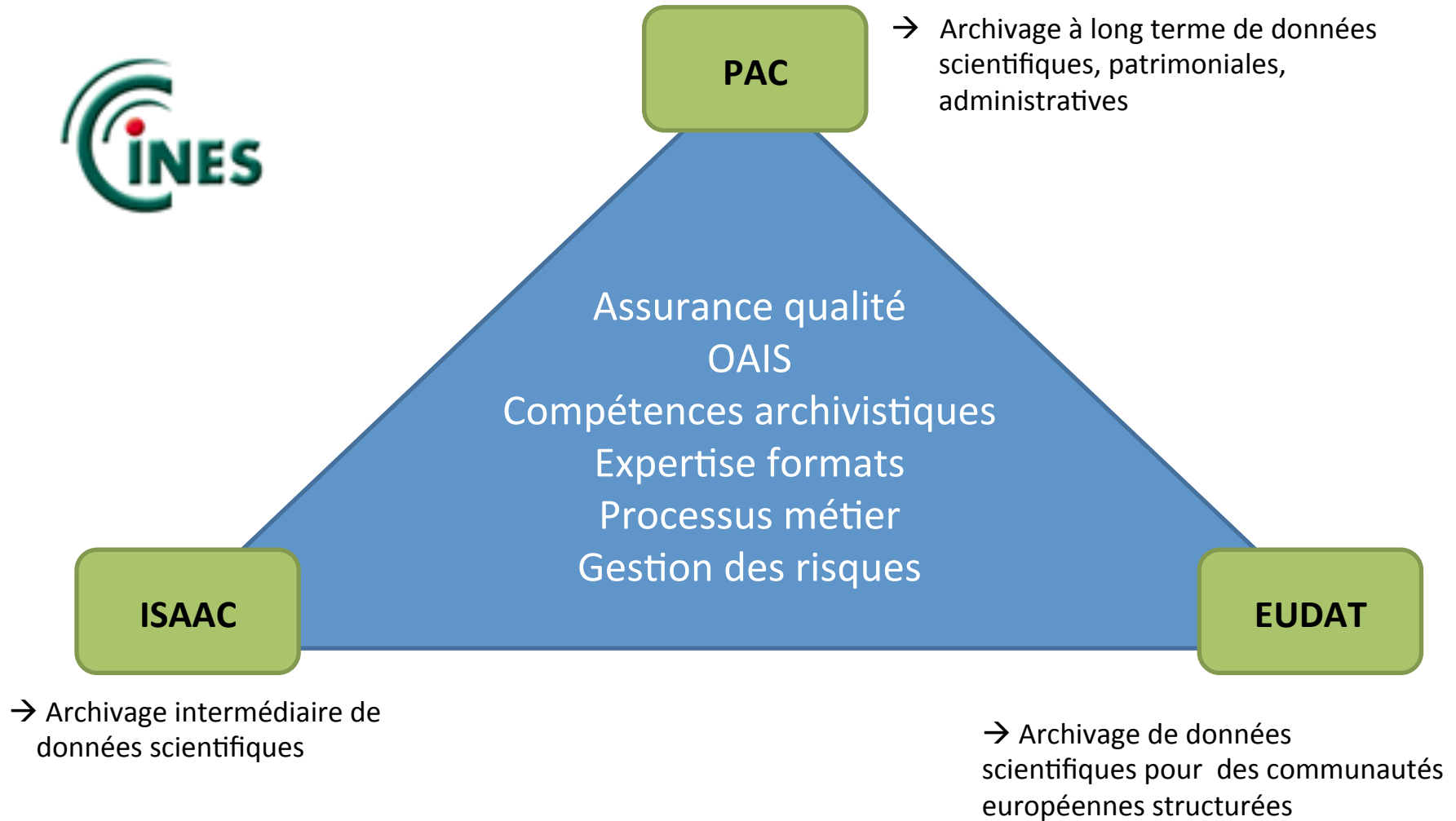
- **Scientific Potential Challenge:** these data sets contain unexploited information, which may give rise to highly useful for joint, multi-disciplinary project.
- **Complexity Challenge:** the data collected by the experimental devices considered in the project is unique and encodes a large typology, well beyond the regular, well-structured data produced in large quantities in the industrial world.
- **Technological et methodological challenge.** The installation of procedures, workflows, algorithms for long term data preservation, as well as the definition of suitable technological frameworks constitute novel investigation domains.

PREDON Consortium

	Volume données	Complexité	Diversification des sources	Structuration au niveau international	Algorithmes et methodologies pour la preservation
IN2P3 HEP	+++	+++	+	++	+
INSU, IRD Astrophysics Earth Sciences	++	++	++	+++	++
CINES INS2I IT, Algorithms, workflows	+	++	+++	+	+++

Example: Archival expertise CINES

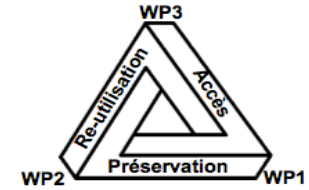
Les services d'archivage au CINES



PREDON: Objectives

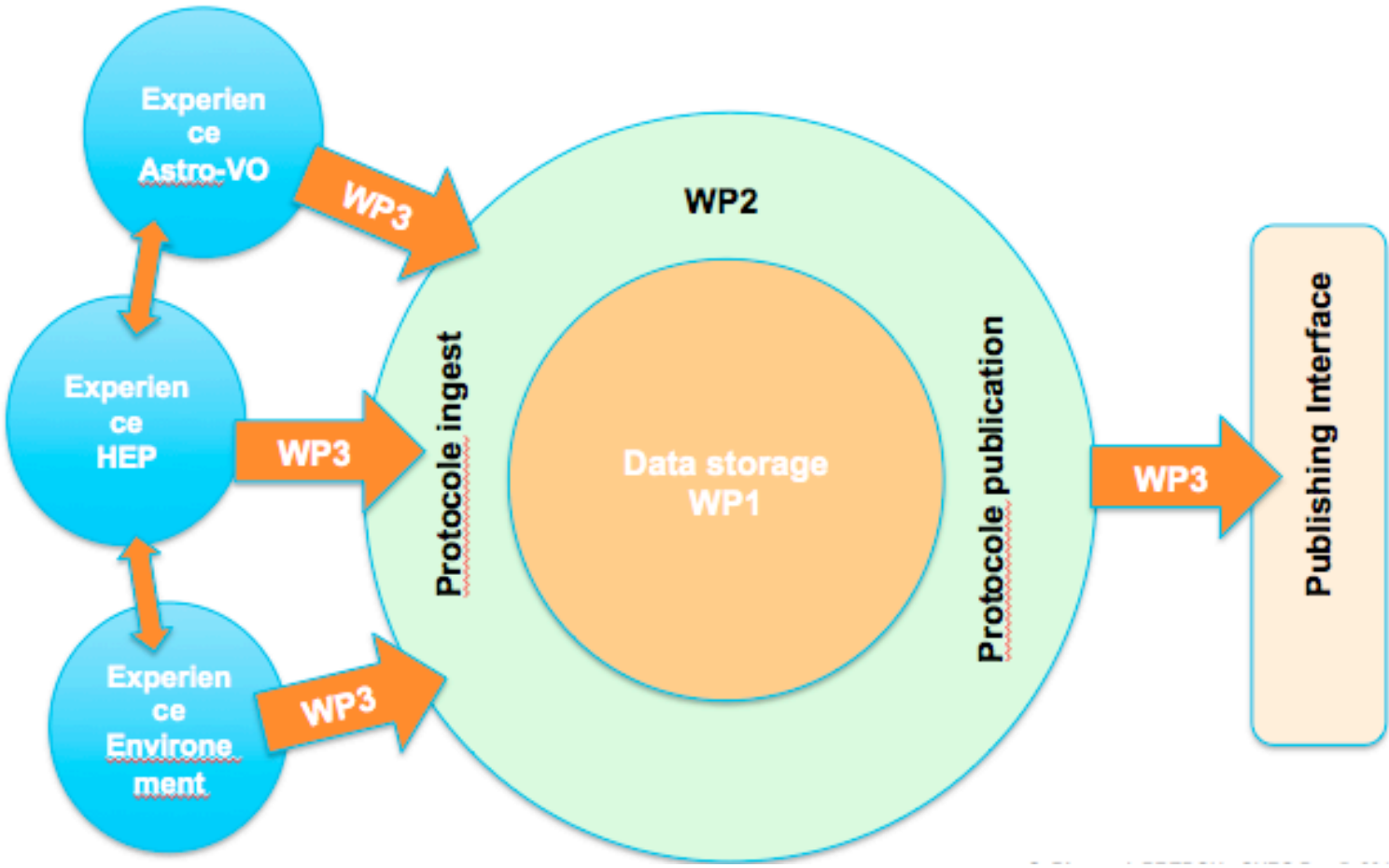
- **Identification** of the scientific and technical **requirements** for an unified approach of data preservation within an multi-disciplinary context.
 - *IOS*: installation of an unified platform to store at long term scientific data in a multidisciplinary context. A demonstrator is proposed within this project.
- Reinforcement of the **coherence and standardisation** of data collection, storage, analysis and access in several scientific domains with complementary needs, leading to a robust and friendly environment for long-term data preservation.
 - *IOS*: installation of a multi-disciplinary mechanism for data preservation standards
- Installation of a **scientific data tracking and supervision system**, such that the information produced during the scientific experiments is followed and centrally tagged at all stages: production, exploitation, archival.
 - *IOS*: definition of a national organisation relative to the scientific data preservation, aimed at supporting main experimental scientific branches producing scientific data towards a traceable long term data preservation
- Reinforcement of **the international cooperation** on this issue in a context of a vast effort to treat large amounts of data sets.
 - *IOS*: permanent links of the consortium with the corresponding international organisations (for instance DPHEP) and the participation of the consortium to the relevant European programs in the field (for instance programs included in the Horizon 2020 agenda).

PREDON as a project



Working Package	Objectives	Participants (*coordinator)
WP1 Technologies and Methodologies	Explore methodologies and technologies suitable for a coherent and robust scientific data preservation in a multi-disciplinary context and on a multi-platform computing centre	CINES* APC
WP2 Algorithms and Workflows	Investigate generic and mathematically robust workflows and algorithms for data mining suited for data and workflow preservation; data- and process-based workflows and mining techniques to be used in a multi-disciplinary environment towards long term data preservation	LAM LIRMM LIPADE* LIPN
WP3 Data formats and interfaces	A parallel approach for data collection, storage, processing, analysis and preservation with the aim to achieve common standards for scientific data treatment	APC CPPM LAM* LPSC
WP4 General coordination	Program coordination, dissemination and international cooperation	CPPM*

Demonstrator



Next Steps

- Workshop Marseille 14/15 November 2013
- Extend discussion to
 - Scientific and technical information (libraries &co.)
 - Legal aspects
 - Economical models
- White paper end 2013 to national funding agencies

Workshop on Data Preservation at ICDE 2014

LOPS@ICDE

WORKSHOP ON LONG TERM PRESERVATION FOR BIG SCIENTIFIC DATA



- <http://lipade.math-info.univ-paris5.fr/lops/>
- LOPS will be held in conjunction with the 30th IEEE International Conference on Data Engineering. Chicago, IL, USA. March 31-April 4, 2014.
- Paper submission deadline November 10, 2013