**Spoke2 - Big Data-Open Data in Life Sciences**

*Table B3: Spoke 2, work packages subjects involved and associated tasks*

|  |  |  |
| --- | --- | --- |
| **WP** | **Tasks** | **Subjects involved** |
| 1. A holistic,  innovative digital architecture for the storage and safe exchange of life sciences big data | 1.1 A novel architecture for big data exchange | UNIMI, ALMAVIVA, TIM |
| 1.2 Governance, management and regulating the access  to services of data in silico |
| 1.2.1 Impatto economico di una raccolta strutturata di big data nelle scienze della vita |  |
| 1.2.2 Pilot Exploiting Genomic Data for Drug Design |  |
| 1.2.3 Fusion of image-tabular data for federated learning of diagnostic models |  |
| 1.2.4 Pilot Digital Strategies for Active Ingredients Synthesis for Pharma |  |
| 1.2.5 Pilot Structure-based in silico target identification |  |
| 1.2.6 Green Radiotherapy |  |
|  |
|  |  |

*WP 1. A holistic, innovative digital architecture for the storage and safe exchange big data for life sciences*

The goal of the WP is to conceive a novel digital infrastructure for the rapid exchange of data related to life sciences by means of machine learning and artificial intelligence, and test them in selected pilot cases.

Task 1.1 A novel architecture for big data exchange – Task 1.1 in the original proposal: a platform will be created for the management of the whole lifecycle of life science data and information, fully respecting data protection regulations.

Task 1.2 Governance, management and regulating the access to services of data in silico – Task 1.2 in the original proposal: this task deals with the governance, management, and regulatory actions related to the access to the services for elaborating the life science data, creating a marketplace of services based on the emerging needs of the research community. Services include i) interfaces for the platforms built for data collection ii) key nodes for collecting and giving value to the data flows from sensors iii) centralised or decentralised models to use and give value to data to calculate relevant inferences for life science. It will develop the strategy for data modeling, as well as for data protection and privacy.

**PILOT CASES: The following tasks will be used to validate the solution proposed in Task 1.1 and Task 1.2.**

Task 1.2.1 Impatto economico di una raccolta strutturata di big data nelle scienze della vita: L'obiettivo di questo task è condurre una valutazione dell’impatto socio-economico dell'infrastruttura di ricerca per l'archiviazione e lo scambio di big data delle scienze della vita sviluppata all’interno dello Spoke 2 del progetto MUSA. Una piattaforma per la raccolta e la condivisione di grandi dati nel ambito del settore life sciences può essere intesa a tutti gli effetti come una infrastruttura di ricerca, ovvero una struttura che fornisce risorse e servizi alle comunità scientifiche per condurre la ricerca e promuovere l'innovazione nei loro campi (European Commission, 2017). In quanto tale, è possibile misurarne l’impatto socio-economico complessivo utilizzando l’analisi costi-benefici, un approccio consolidato per la valutazione degli investimenti, e che recentemente è stata utilizzata da membri del DEMM anche per la valutazione di progetti di Ricerca e Sviluppo, in particolari collegati alle grandi infrastrutture di ricerca (Florio, 2019). L’utilizzo della analisi costi-benefici per la valutazione sociale delle infrastrutture di ricerca richiede di considerarne le specificità in termini di benefici generati. Questi possono sono collegati ai benefici diretti ed indiretti ottenuti dagli utilizzatori dei servizi dell’infrastruttura: è il caso ad esempio del valore per gli scienziati delle pubblicazioni prodotte grazie all’infrastruttura, o delle ricadute positive di cui possono avvantaggiarsi le imprese che contribuiscono alla realizzazione della stessa. A questi si accompagnano i benefici per l’utilizzo futuro della conoscenza (il più delle volte di difficile misurazione ex ante) e il valore intrinseco della conoscenza come bene pubblico, ovvero il valore sociale attribuito alle scoperte scientifiche dalla collettività.

Task 1.2.2 Pilot Exploiting Genomic Data for Drug Design: Creazione di una Virtual Biobank con dati derivati da profiling di RNA sequencying di modelli sperimentali umanizzati. Validazione di strategie di “Secure Transmission” dai sequenziatori next generation sequecying (NGS) e loro digitalizzazione. Applicazione di modelli di Inteligenza artificiale per analisi dei dati.

Task 1.2.3 Fusion of image-tabular data for federated learning of diagnostic models: Creazione di un approccio ripetibile per studi diagnostici multi-centrici, basati su federated learning e dati di tipi eterogenei. Predisposizione di banche dati locali e standardizzazione dei metadati. Predisposizione e sviluppo di modelli d’Intelligenza Artificiale privacy-aware, caratterizzati da alto throughput locale e scambio di parametri in modalità federated learning. Il caso di studio di riferimento si focalizza sulle tecniche di fusione di informazioni cliniche multiple raccolte da centri indipendenti per la diagnosi precoce e la stadiazione del cancro alla prostata. Il pilot stabilirà una pipeline di AI ripetibile che includa 1) raccolta di dati fortemente eterogenei (tipo immagine e di tipo tabellare) 2) digitalizzazione e fusione al fine di (i) massimizzare il l'accuratezza della diagnosi precoce (ii) valutare il suo beneficio differenziale rispetto alla sola imaging fornendo una stima affidabile del guadagno in precisione, rispetto al costo della raccolta aggiuntiva e della pre-elaborazione.

Task 1.2.4 Pilot Digital Strategies for Active Ingredients Synthesis for Pharma: Digital and technology-driven strategies for the development of innovative and sustainable synthesis of active ingredients and crucial intermediates, including antivirals, according to the green chemistry and circular economy principles.

Task 1.2.5 Pilot Structure-based in silico target identification: Il task si prefigge di sviluppare un metodo structure-based di target identification attraverso metodi ottimizzati di inverse docking. In altri termini, il metodo si prefigge di identificare con quali proteine di interesse terapeutico (target) un dato composto possa interagire. Tale metodo verrà implementato in un webservice, che consenta un utilizzo semplice intuitivo e distribuito di tale risorsa. L’identificazione dei probabili target implicati nell’azione di una data molecola può trovare numerose e notevoli applicazioni sia nell’interpretazione di saggi fenotipici sia nella razionalizzazione dell’attività terapeutica di estratti naturali.

Task 1.2.6 Green Radiotherapy: NEED DESCRIPTION.

*Table C.5: Targets and Milestones for Spoke2 activities.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Type** | **No.** | **Short description** | **Means of verification/KPI** | **Month** |
| WP1: A holistic, innovative digital architecture for the storage and safe exchange of life sciences big data | | | | |
| Targets | T2.1 | Platform for the management of the lifecycle of life science data and information | N. pilot executions on the platform | 24, 30 |
| T2.2 | Marketplace of services based on the emerging needs of the research community in life science | N. services available  N. access to services | 36 |
|  | MS2.1A | Definition and adoption of a protocol for data management | N. protocol instances | 18, 30 |
|  | MS2.1B | Platform and marketplace | N. total executions on the platform | 24, 36 |