Abstract—This paper introduces a new information technology: ma³tch (autonomous anonymous analysis). Ma³tch enables virtual information integration to build a ‘dynamic networked collective intelligence’ without infringing upon security, confidentiality, privacy and/or data protection regulations. It provides organizations with information and knowledge advantages.

The ma³tch technology is empowered by a decentralized information oriented architecture: a ‘privacy by design’ framework that uses distributed agents to facilitate decentralized but integrated information access, processing and analysis. It shapes a ‘virtual information cloud’ between autonomous organizations that enables secure, integral and intelligent real time information analysis. Relevant information and knowledge distributed between autonomous organizations is automatically detected and applied throughout the network as soon as it emerges.

The dynamic design principles allow practically any type of (cross domain) information to be virtually integrated: government, commercial, intelligence, law enforcement, financial, telecom, biomedical, compliance, etc., without infringing privacy, confidentiality, security or data protection rules and regulations. It advances both privacy AND knowledge beyond conventional limitations.

Keywords—decentralized information oriented architecture; privacy by design strategy; virtual information cloud; dynamic networked collective intelligence; autonomous anonymous analysis

I. INTRODUCTION

For executives and policy makers information technology is like a double edged sword. Collecting, combining and analyzing information offer huge benefits for business and governments. But at the same time when personal data are involved that information easily infringes upon the privacy of citizens. Balancing between privacy and knowledge is a continuous struggle.

Centralizing and combining data have many advantages: increased knowledge and insight, diagnoses, identification of trends and threats, strategic and competitive advantage, etc. But it also leads to severe risks of unlawful access, hacking, identity theft, fraud, manipulation, exclusion from insurance, etc. As the amount of combined information grows, so do the risks and threats. ‘In the 21st Century, bits and bytes can be as threatening as bullets and bombs,’ said Deputy Defense Secretary William J. Lynn III during a speech after a cyber-attack at the Pentagon [1]. It emphasizes both the power and threat of information and the importance to protect that information [2], [3].

When information is distributed between (multinational) organizations and/or governments many implementation barriers emerge (business, organization, information and technology). These barriers significantly limit the information parties are allowed, willing and able to combine. Even at a national level it is often impossible to combine and analyze distributed government databases. What about international collaboration where national interests and (conflicting) national legislation come into play?

This paper introduces the ma³tch technology and the ‘privacy by design’ framework used by EU Member State Financial Intelligence Units, that mitigate these barriers. They enable virtual integration and real time analysis of distributed information, knowledge and technology capabilities, without the need for parties to move sensitive information beyond their own premises. Although the technology is designed for intelligence purposes, the dynamic architecture makes it suitable for many other industries and cross domain integration: government, commercial, financial, telecom, biomedical, telecom, law enforcement, compliance, etc.

II. PARADOX: PRIVACY AND KNOWLEDGE

Financial Intelligence Units (FIUs) are central, national intelligence agencies responsible for receiving, analysing and disseminating disclosures of financial information to the competent authorities (e.g., law enforcement or prosecutorial authorities) in order to combat money laundering and terrorism financing.
Depending on the type of FIU [4] and national legislation, typically FIUs receive millions of financial transactions, from thousands of reporting institutions (e.g., banks, money or value transfer providers, dealers in high-value goods, casinos, etc.), each with their own industry specific data domain and structures. In addition FIUs have direct or indirect access to dozens of (heterogeneous) national government, commercial and public information resources (e.g., criminal investigations, convictions, chambers of commerce, bank accounts, etc.).

In the analysis process FIUs face a paradox: FIUs must exchange ‘need to know’ information with national and international partners to bring relevant information together, while guaranteeing strong and effective protection of their privacy and personal data [5].

During this pre-investigative intelligence phase, FIUs can exchange information with each other as long as the information does not leave the FIU premises. Only suspicious information is disseminated to law enforcement and only after prior consent is given by the FIU that disclosed the information. This allows FIUs to cooperate effectively and efficiently without risking FIU intelligence ending up in police systems. Using the same mechanism FIUs can include national government, commercial and public information sources.

FIU.NET is a sophisticated decentralized computer system, funded by the European Commission since 2004, that enables and enhances the information access, analysis and exchange between EU Member State FIUs. It shapes a virtual information cloud between the FIUs and their (over 550) distributed government, commercial and public information resources and it enables real time analysis of distributed dynamic information and knowledge otherwise legally, organisationally, technically, and/or financially impossible to achieve.

FIU.NET is based on 3 core design principles: autonomy, decentralization and flexibility. Autonomy guarantees information owners full governance over information they connect to the system, and that no other party (or service provider) has the possibility to gain access to that information without consent of the information owner. It ensures information owners are never forced to comply with externally imposed data standards, rules regulations or limitations. The autonomy principle is crucial to maximize the amount of information that parties are allowed, willing and able to connect.

Decentralisation guarantees autonomous information control and it physically enforces ‘privacy by design’ regarding storage, processing, analysis and exchange. Information is only physically stored and processed at the premises of the information owner, and is optionally (controlled by the owner) distributed where and when access to that information is needed. This means there is no need for remote access (and risks) to the information. Decentralization also enhances scalability, speed (local access), availability and robustness (there is no central point of failure).

Flexibility enables autonomous (data) governance. Instead of having to comply with imposed data standards, rules and regulations of a central authority, ‘flexibility by design’ allows each party to integrate their data as is, and on its own terms. Data governance (standards, rules, regulations, capabilities) can be dynamically adjusted and locally enforced by each information owner. Although enforcing a single standard is much easier, the specific information position of FIUs makes this virtually impossible, unless information quality, quantity, speed and capabilities are severely limited and reduced.

The decentralized information oriented architecture (Fig. 1) solves this by shaping a virtual enterprise architecture [6]. It directly connects data owners (organisations) without 3rd party involvement. Custom data from the back office are virtualized by stores into dynamic structured virtual data elements. All (distributed) data elements can be hierarchically linked and are governed by their information owner using dynamically configurable information resources. Resources specify custom business rules (dynamic data structures, data retention time, prior consent, information handling, etc) for each of the local information sources the FIU has access to and negotiate bilateral or multilateral capabilities. This enables both manual (user driven) and automated (agent driven) iterative case building and real time integral analysis of decentralized information and knowledge. Like Lego blocks, the virtual architecture can be dynamically adjusted to the legal, organisational information and technology needs of each party.

![Decentralized Information Oriented Architecture](image)

The ‘privacy by design’ decentralized information oriented architecture builds a virtual information cloud between all parties connected to the network and enables parties to quickly adapt to emerging changes and opportunities. The resulting
virtual platform bridges and harmonizes legal, organizational, informational and technological differences in a complex system [7] of autonomously governed parties. It establishes a dynamic information architecture and virtual enterprise architecture without infringing upon local governance, privacy, security and confidentiality.

In many ways it is the opposite of conventional information architectures and solutions. Instead of enforcing standards, standards are dynamic and virtualized. Instead of relational data storage, data are stored hierarchically. Unlike federated databases [8] there is no remote access or (costly) heterogeneous implementation. Compared to service oriented architectures [9] where information moves between services, information remains where it is. Instead, services (agents) are moved to the information. Compared to conventional analysis where information is first combined and then analyzed, information is first analyzed and only relevant information is combined by pushing or pulling information. Unlike cloud solutions, there is complete control where information is stored and processed. Instead of storing and processing information deep ‘inside’ boundaries of a service provider [10], the information architecture is reversed: in a sense it is a cloud inside out! The physical (vertical) decentralized processes flow of the information oriented architecture consists of 5 (horizontal) virtually integrated layers. Fig. 2 visualizes a generic information flow from an information provider (A) to one or more information consumers (B, C):

At both parties the information flows through all layers:

1. The data access layer connects any kind of data or service. Parties can connect local information ‘as is’ to the virtual information cloud (reduces time, requisites and costs).
2. The data virtualization layer virtualizes diverging data domains or services into uniform (standardized or generic) virtual data elements (harmonizes information).
3. The processing layer virtually integrates information with other parties, but also provides standardized local data processing and analysis capabilities (harmonizes technology).
4. The governance layer guarantees local autonomy and maximizes information that parties are able, allowed and willing to contribute (harmonizes governance and processes).
5. The exchange layer directly connects parties (direct bilateral connections guarantee that there is no intermediary data storage or third party that imposes or limits capabilities).

Since most information consumers are also information providers, similar information flows emerge in opposite direction. Sensitive information remains decentralized under full (physical and logical) control of each information owner. This guarantees that other parties cannot access the information but does allow information and capabilities to be virtually integrated and analyzed (when and where allowed) to identify relevant ‘need to know’ information and knowledge. Information is only physically stored and accessible where and when access to that information is needed and allowed.

---

**Fig. 2. Decentralized Information Oriented Architecture Processes**

- **Access Layer 1:** connects any kind of service or data domain, source and structure.
- **Virtual Layer 2:** uniform hierarchic virtual data elements: harmonizes information.
- **Processing Layer 3:** standardized local data processing and analysis: harmonizes technology.
- **Governance Layer 4:** local governance: guarantees autonomy: maximizes information.
- **Exchange Layer 5:** direct connections: no central authority that enforces/limits capabilities.

---

[25]
Ma³tch (autonomous anonymous analysis) is the generic analysis process to accomplish virtual information integration: the ability to combine information, knowledge and capabilities from multiple distributed organizations without bringing the data physically together. It is based upon the same core ‘privacy by design’ principles of the information oriented architecture: autonomy, decentralization and flexibility.

Ma³tch virtually integrates information and/or knowledge using distributed agents: programs with small and well-defined (analysis) tasks that run distributed in the processing layer. Local agents transform information and personal data into anonymized filters that can be (selectively) shared: the filters only contain characteristics of the original data, not the actual data itself. At their destination agents re-integrate the characteristics of the anonymous filters with local information resources. Ma³tch agents can be used for any kind of local, shared or distributed integral analysis (operational, strategic, social networks, geo-tagging, prediction, correlation, entity extraction, etc.).

The virtual integration process (Fig. 3) derives knowledge from the local resource (induction). The resulting knowledge is used locally or is distributed and integrated with remote resources (deduction). The ma³tch process identifies relevant information and knowledge links between all parties in real time. Even though no sensitive information is exchanged, all organizations can identify where and when other relevant information is available. Each organization increases its information position when it connects new information, resulting in a dynamic voluntary [11] and reciprocital [12] [13] knowledge network.

Ma³tch achieves cross-organization information and knowledge pull [13]: the ability to find, access and attract information, knowledge and resources that are relevant and valuable, even if currently all parties are unaware that the information exists. The process involves 3 steps: autonomous (access), anonymous (attract), analysis (achieve).

1. Autonomous (access)

Autonomy is guaranteed by the information oriented information architecture, and maximizes the information parties are allowed, willing and able to connect. Each information owner controls what data are included in the filter, how long the filter is valid, what the precision of the filter is, with which parties the filter is shared, and after a hit if, when and what personal data are exchanged. The fact that information is only stored when and where access to that information is needed maximizes proportionality and subsidiarity [14:15], but also empowers availability by enabling (virtual) information access.

2. Anonymous (attract)

Anonymization is non-trivial. Contrary to encryption, anonymization must be irreversible. Privacy sensitive data is considered anonymized when it cannot be directly or indirectly connected to an identified or identifiable individual [15]. Pseudonimization, for example using hashing values, does not suffice, as it still may be used to identify an individual. But even when all personal data is replaced with random numbers, or is completely removed, the remaining data may still be used to link it back to an individual [16]. To guarantee anonymity, individual dimensions and records are minimized and aggregated in such way that it is impossible to distinguish or link to individual personal records. Only characteristics, captured in the anonymous filters are shared with selected parties to ‘magnetize’ and attract relevant information.

3. Analysis (achieve)

Analysis integrates received anonymous filters with local information. This provides each organization with unique operational, tactical and strategic knowledge products integrated with their own sensitive information. Local information in each organization is ‘tagged’ in real time with the virtual emerging information and knowledge products: where other relevant information can be found, trends, threats, risks, etc. This analysis process requires flexibility (agents) to allow continuous dynamic (iterative) adjustment to emerging results and achieve integrated results.

The ma³tch process virtualizes analysis existing techniques to ma³tch: information, knowledge and technology.
IV. MATCH EXAMPLES

Full technical details of the analysis methods are out of scope, but basic principles and processes are briefly discussed.

1. Information Match

Examples of anonymization, aggregation and comparison algorithms that can be used to match distributed information are: fuzzy logic, hash tables, Bloom filters, transliteration, n-grams, and approximation techniques. The algorithm used in FIU.NET was initially designed for the Intersect project [17] for real time concept extraction and text analysis. It allows fast anonymous distributed cross matching: a million records easily fit in a 1 MB anonymous data structure and can be matched in less than 0.3 milliseconds on a single 3Mhz processor core, making it possible to check hundreds of distributed information resources in real time. As simplified example, an information resource contains:

- Philip Tattaglia (12/28/16)
- Luka Brasi (3/13/26)
- Johnny Fontane (10/7/27)

The anonymization algorithm minimizes these 3 individual records into a single combined anonymous 4 character fuzzy logic data structure: ‘tnUG’. This 4 character code captures the ‘characteristics’ of the combined original sensitive information, making it impossible to recover the individual records. The extreme data minimization enables (configurable) false positives (collisions) that further enhance anonymity. In addition, the information owner controls which data are included in filter, and if, when and where filters are shared (multiple filters can be created for a single dataset, for example with lower accuracy for sensitive data). Other parties that receive the filter can use it to match local sensitive data against the anonymized data structure ‘tnUG’ without knowing the underlying data. The example code ‘tnUG’ allows matching of transliterations, permutations and approximations with 99% accuracy: all following examples match positively with the data structure, even though they use different punctuation marks, date formats, and/or contain spelling errors. Positive hits are optionally or automatically followed up for (anonymous) validation, compliance check, and/or a fully detailed ‘need to know’ information exchange:

- Fillip Tataglia (12/28/16)
- Tattaglia Philip (28-12-1916)
- BRASI, 13-03-1926 (Luca)
- Luka Anthonius Braazzi (3-13-1926)
- Jonnie Fontarië (07-10-1927)
- Джони Фонтане (19270710), etc

Fig. 4 visualizes the process where organization A generates an autonomous match filter. Only the characters ‘tnUG’ are exchanged. Organization B locally checks its sensitive information against the anonymous filter ‘tnUG’ to identify what local information matches with organization A: Information links between the organizations are identified in real time as soon as they emerge.

2. Knowledge Match

Besides virtual information integration, match technology also allows virtual knowledge integration. Knowledge discovery (data mining) is the process of discovering patterns from large data sets using statistics and artificial intelligence. Intelligent data analysis can identify, describe and predict trends, threats, profiles, behaviors, etc. Existing data mining techniques can be extremely powerful in combination with the match: it enables global networked knowledge discovery and distribution. An example of a predictive data mining technology is a neural network. Neural networks capture and apply knowledge hidden within data. Like the human brain, they learn from experience. The more often a certain combination leads to a specific outcome, the stronger the associations and certainties get (simplified example in Fig. 5).

Often neural networks are biased by internal processing logic, data quality and noise of each organization. As a result many predictions are unreliable, and if they are correct they ‘explain’ what the organization already implicitly knows (or thinks it knows).

Match allows parties to integrate external knowledge with
their own local information, and identify which data potentially have a higher risk based on that external knowledge. Existing local data are labeled with external knowledge tags. The ability to combine the results of multiple neural networks from multiple organizations in real time, leads to far more reliable and useful knowledge products.

Fig. 6. Neural network match example between 2 organizations.

Ma’tech virtually integrates the emerging collective knowledge between organizations. Even better, organizations not only learn and apply what they already implicitly know, but they learn and apply the distributed collective knowledge: a high risk transaction is identified (Fig. 6) because transactions with similar patterns are identified in 5 other organizations as high risk. This transaction would not have been identified based on local knowledge only. Similarly other data mining technique can ma’tech (e.g., clustering, classification, decision trees, n-grams, etc). Prototyped examples include prediction of the gender or nationality of a person, based on first or last names of a person. All organizations locally train the ma’tech using their own local data, and utilize the combined networked intelligence to ‘predict’ missing information. There are countless other examples, but the essence is to jointly build, share and utilize common standardized knowledge without the need to disseminate sensitive information.

3. Strategic Ma’tech

Statistical ma’tech analysis enables strategic decision making. Statistics are automatically (locally) extracted from sensitive information, and virtually integrated to achieve integral knowledge products. Ma’tech analysis of money flows identifies hidden distributed financial trends in real time, for example to provide insight in local (known) versus global (unknown) money flows (Fig. 7). It reveals money flows that are locally not available and otherwise would have remained undetected (i.e., FI and PL are unaware of the money flows from NL, NL is unaware of the money flow to BG).

Fig. 7. Local money flows versus global money flows (perspective NL)

Trends, threats and irregularities can be detected as soon as they emerge, and automatically integrated with local data. For decision making this real time global view provides huge strategic advantages to all organizations participating in the ma’tech analysis. Combining the results of strategic ma’teches with directed neural network ma’teches or anonymous ma’teches, or even the underlying sensitive data, provides operational and knowledge insight into these strategic results.

4. Social Network Ma’tech

Social network ma’tech enhances social network analysis by creating knowledge about missing entities and links. Social network analysis maps and measures relationships and flows between people, groups, organizations and other connected information or knowledge entities. Common measures that provide insight into the various roles and groupings in a social network include: degree (number of direct connections), betweenness (number of indirect connections), closeness (distance to entities in network) and bridges (connecting different networks). But a single piece of missing information severely impacts these measures. Virtual integration (combining the information-, knowledge- and strategic ma’tech), identifies new and/or similar entities, links and relations between (otherwise disconnected) networks (Fig. 8).

Fig. 8. Example where ma’tech completes ‘loose ends’.

It allows parties to identify information links with other information (distributed across databases) to determine which role that information has in the ‘global’ network. Parties can even identify that two or more entities that locally seem to be disconnected, are actually related to each other based on information available at the other party (missing links).
5. **Technology Ma’tch**

Ma’tch minimizes financial, legal, organizational, technical and informational implementation barriers, and maximizes the information and knowledge position of all parties connected to the virtual cloud. As connected parties share a common virtual enterprise architecture (and virtual information architecture), they can jointly develop or acquire new (commercial) information sources and services cost effectively, and seamlessly integrate these with their internal systems and databases supporting privacy, security or confidentiality. It reduces individual costs and further enhances and harmonizes their capabilities. It enables connected parties to act and operate as a single virtual enterprise.

This enhances the local technology capabilities of the parties involved, and enables access to information and technologies otherwise too difficult or too costly to achieve. Examples are (anonymous) access and integration with (commercial) databases, but also with generic services like entity-extraction, geo-tagging, content translation, rule engines, etc.

VI. **VIRTUAL INFORMATION CLOUD INFRASTRUCTURE**

The physical infrastructure of the information oriented architecture and ma’tch is dynamic as well, and can run on low-end servers, and is vendor independent. In principle multiple nodes can run on a single (virtual) server. Each time privacy, security, or information access is a concern, a new physical distributed node can be added. Organizations can directly or indirectly connect to each other (Fig. 9). Organizations may share a physical server (Organization A and B), have a dedicated server (Organization C), or use delegated servers for each of their resources (Organization D).

![Fig. 9. Virtual Information Cloud Infrastructure](image)

With ma’tch there is no need to exchange bulk information to achieve integral data fusion and analysis between parties, and there is no need to entrust sensitive data to a third party. Timeliness increases as relevant information is identified in real time, and is pushed or pulled when needed: no unnecessary information needs to be processed, cleaned or reviewed, which also prevents huge backlogs. Quality and reliability are enhanced as information can be processed, interpreted, amended and checked for compliance at the source.

Although the technology is designed for intelligence purposes, the dynamic architecture (due to the large variety of information FIUs process) makes it suitable for many other industries, and cross domain integration: government, commercial, financial, telecom, biomedical, telecom, law enforcement, compliance, etc.

At a national level many distributed government agencies deal with privacy sensitive information. They are expected to work together but are often (not without reason) limited by privacy regulations. This may for example prevent or limit linking weapon permits to mentally ill patients, detecting abuse of social security systems, linking sexual offences to social services. When incidents happen, in many cases the information was available, but not connected where and when it was needed. Ma’tch makes it possible to find where this (otherwise unknown) relevant information is available in real time. This can help social security or tax officials to identify fraud, hospitals to identify emerging health threats, banks to conduct joint compliance and risk assessments, etc. It can be used to analyze distributed telecom/internet traffic to combat cybercrime, and could provide an alternative to centralized DNA or electronic patient file databases; at a national level but also at an international level.

At international level benefits even further increase. Countries have lots of comparable agencies that perform similar but distributed activities in each jurisdiction. Ma’tch allows agencies of countries with a shared mission to identify relevant remote information, and use each other’s knowledge to learn, cooperate and collaborate. For example to anonymously ‘match’ convictions between countries in order to identify criminal activities moving across borders. When privacy sensitive data are involved, for example in international collaboration between intelligence agencies on money laundering and terrorist financing, ma’tch identifies links, threats, trends and shifts between countries that otherwise remain undetected. Similarly hospitals could ma’tch biomedical data to detect trends and threats in healthcare in real time. Ma’tch shifts the focus from reactive to proactive.
International law enforcement agencies (like Interpol and Europol) may use ma³ch for real time identification of relevant (need to know) information on ongoing cases. It provides virtual access to information when and where it is needed (financial, telecom, intelligence, etc.) without the need for parties to expose sensitive information. The dynamic architecture allows matching virtually any kind of information from any source, at minimal costs. Also cooperation between Europe and the United States [18] (like PNR, TFTP) could benefit from the ma³ch, and build a collective intelligence and knowledge to identify relevant information, trends and threats as soon as they emerge, without infringing upon security, confidentiality and privacy.

Ma³ch and the information oriented architecture empower data minimization, proportionality and subsidiarity: the guiding principle within the European Union that matters ought to be handled by the smallest, lowest or least centralized competent authority [14]. It also enables the availability principle and the objective of the European Union for Member States to obtain information rapidly from private or public authorities of another Member State without the use of coercive measures. Despite of different legal systems and national regulations ma³ch enables collaboration opportunities far beyond existing limits and barriers.

The information oriented architecture has proven its value as the foundation for FIU.NET: it enables autonomous governed information access, exchange and analysis between the national FIUs (Financial Intelligence Units) of the EU member states. FIUs can virtually integrate their information, without the need to place sensitive information beyond their organization’s premises. Between the EU FIUs there are over 550 information resources, each with their own information standards, policies, processes and data protection rules and regulations. It provides FIUs with secure (virtual) access to government and commercial databases through a customized unified interface, and allows them to innovate and acquire new services and information resources as one virtual enterprise.

With the information ma³ch FIUs can detect relevant information between each other, but also with commercial data sources, without leaving any traces. Within this diverse and complex environment, ma³ch technology enables joint information analysis and innovation capabilities otherwise impossible to achieve. But also at a national level FIUs can benefit from ma³ch to enhance local analysis capabilities, and by using it for (cross border) reporting by exchanging ma³ch filters with reporting institutions (reducing administrative burdens, and enhancing national investigations).

FIUs have already named the emerging information and knowledge ‘carefully powerful’, and now face a new challenge: to deal with the emerging streams of real time information and knowledge.

Ma³ch builds a dynamic networked collective intelligence (Fig. 10) that provides participating parties with tremendous information and knowledge advantages. Relevant information links and knowledge distributed between parties are detected in real time as soon as they emerge. It gives governmental and commercial organizations a unique information and knowledge position that otherwise would be impossible to achieve. It provides a global overview for strategic decision making and generates knowledge that otherwise would remain distributed, hidden and undetected. Integral operational and strategic analysis on distributed data without infringing upon privacy or security advances results beyond what is otherwise possible. Privacy and knowledge do not conflict but strengthen each other, maximizing both privacy AND knowledge.

The information oriented architecture provides a dynamic and flexible foundation for decentralized information access, exchange, and analysis that enables unique strategic advantages. It enhances data minimization, proportionality and subsidiarity and maximizes voluntary and reciprocal communication, coordination, cooperation and collaboration. It strengthens coherence and bridges legal, organizational, informational and technological differences between autonomous parties. It shapes a dynamic virtual enterprise that enables parties to operate as one, and quickly and cost effectively adapt to emerging changes and opportunities.

Understanding the ma³ch technology requires a paradigm shift, not just in understanding what is possible, but perhaps even more important in how organizations, governments, central authorities, legislators, data protection offices and information managers look at collaboration and information technology problems and solutions. Information from any national and/or international data source can be virtually integrated. It advances both privacy AND knowledge beyond conventional limitations: ma³ch opportunities start where other possibilities stop.


